EVOLUTIONARY GENOMICS OF

METHYL-ACCEPTING CHEMOTAXIS PROTEINS

A Dissertation
Presented to
The Academic Faculty

by

Roger Alexander

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Biology

Georgia Institute of Technology
December, 2007

EVOLUTIONARY GENOMICS OF

METHYL-ACCEPTING CHEMOTAXIS PROTEINS

Approved by:

Dr. Igor Zhulin, Advisor
Joint Institute for Computational Sciences
University of Tennessee
Oak Ridge National Laboratory

Dr. Eberhard Voit
School of Biomedical Engineering
Georgia Institute of Technology

Dr. Stephen Harvey
School of Biology
Georgia Institute of Technology

Dr. I. King Jordan
School of Biology
Georgia Institute of Technology

Dr. Nick Hud
School of Chemistry and Biochemistry
Georgia Institute of Technology

Date Approved: 6 September, 2007

For Keren and Mica, Mom and Dad, and family, with much love.

Mica, wait until you hear the story of George the giant mosquito.
Bugs as big as buses, enzymes the size of peas --
it’s pretty cool.

Eat your vegetables!

ACKNOWLEDGEMENTS

I would like to thank Ron Kurti for help with cascading style sheets, Chris Rao for discussions about chemotaxis in *Helicobacter pylori*, Brian Crane for discussions about the structure of the chemoreceptor complex in *Thermotoga maritima*, Phil Aldridge for discussions of flagellar regulation and evolution, Ann Stock for her perspective on the molecular basis of feedback in chemotaxis, and Uri Alon and the participants in the Kahn-Minerva winter school for providing a stimulating forum that strengthened my understanding of biological design principles. I would like to thank my thesis committee members for their time and insight.

I really enjoyed the three (four?) years I spent thinking deeply about chemotaxis with Kristin Wuichet, bouncing ideas off of each other and generally motivating each other to figure things out in a fun upward spiral. I learned a lot from Luke Ulrich about Linux and computer programming; I regret only that our metaphysical differences kept us from making more progress together in the understanding of evolution. I want to apologize to Siddarth Joshi for stealing his project and thank Igor Zhulin for handing it off to me. Chemotaxis has proved to be a fascinating entry into the world of computational biology. I learned a lot from Igor about what kinds of questions to ask and how to think strategically about the best approaches for answering them. The postcard on his office door of a still life with wine, bread, and cheese says it all. I have just one more question and then I will stop.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

Alt	Alternate output
AS	Amphipathic Sequence
BLAST	Basic Local Alignment Search Tool
BLOSUM	Block Sum
CCW	counter-clockwise
CDD	Conserved Domain Database
Cheops	Chemotaxis Operons
CheY~P	phosphorylated CheY
[CheY~P]	concentration of phosphorylated CheY
COG	Cluster of Orthologous Groups
CRC	Cyclic Redundancy Check
CW	clockwise
DDBJ	DNA Database of Japan
DNA	Deoxyribonucleic Acid
EMBL	European Molecular Biology Laboratory
F	Flagellar
FBS	Flexible Bundle Subdomain
FRET	Fluorescence Resonance Energy Transfer
Glx	glutamate or glutamine
HMM	Hidden Markov Model
Hpt	Histidine phosphotransfer
IC	Information Content

MCP	Methyl-accepting Chemotaxis Protein
MCP_CD	Methyl-accepting Chemotaxis Protein Cytoplasmic Domain
MEGA	Molecular Evolutionary Genetics Analysis
MH	Methylation Helix
MiST	Microbial Signal Transduction
ML	Maximum Likelihood
MSA	Multiple Sequence Alignment
NCBI	National Center for Biotechnology Information
NJ	Neighbor Joining
NMR	Nuclear Magnetic Resonance
NW	Needleman-Wunsch
OCS	One-Component System
PAM	Point Accepted Mutation
PDB	Protein Data Bank
Pfam	Protein families
PSI-BLAST	Position-Specific Iterative BLAST
PSSM	Position-Specific Scoring Matrix
r.m.s.d.	root mean-square deviation
RDBMS	Relational Database Management System
Refseq	Reference Sequences
SMART	Simple Modular Architecture Research Tool
SQL	Structured Query Language
SW	Smith-Waterman
TCS	Two-Component System
Tfp	Type-IV Pili

TM	Trans-Membrane
UC	Unaligned Cytoplasmic
UM	Unaligned Membrane-bound

SUMMARY

The goal of the Zhulin lab is to gain a better understanding of signal transduction mechanisms in prokaryotes using the tools of computational biology. The specific focus of this research project was to characterize the cytoplasmic domain of methyl-accepting chemotaxis proteins (MCP_CD), a protein domain central to the function of the most complex signaling network found in prokaryotes. The chemotaxis signal transduction network enables cells to sense and respond to multiple external and internal cues by actively navigating to a more optimal environment. Prokaryotes are too small to sense gradients in space; instead, they sense gradients in time using a feedback loop in the chemotaxis protein network that gives the cell a memory of recently experienced stimuli. MCP_CD is a central part of this memory circuit, but the coiled coil structure of the domain made it difficult for traditional tools of computational biology to analyze. The central goal of this research project, then, was to develop a new method for analysis of the domain, and then to gain new insight into its function and evolution. In the process, two other significant contributions were made that built on work by other members of the lab.

Research advance 1: Characterization of the MCP_CD protein domain.

Before this research project, MCP_CD was known to have two distinct functional regions: (1) the signaling region where activating and de-activating interactions with the

histidine kinase CheA take place, and (2) the methylation region where adaptation enzymes CheB and CheR store information about recently experienced stimuli.

The result of this project is the classification of almost 2000 MCP_CDs into twelve subfamilies. The unique mechanism of evolution of the domain has been clarified and precise boundaries of the adaptation and signaling regions determined. A new functional region, the flexible bundle subdomain (FBS), was identified and its contribution to the chemotaxis signaling mechanism elucidated by analysis of conserved sequence features. The characteristic pattern of coiled coil knob residues in the FBS seems to act as a conduit for signals from the diverse sensory domains found in MCPs, and to cause the MCP coiled coil to bend in response. Conserved and variable sequence features in the adaptation and signaling subdomains led to a better understanding of the evolutionary history of the adaptation mechanism and of alternative higher-order arrangements of the receptors within the membrane.

Research advance 2: Development of a sensor / kinase correlation algorithm to couple diverse MCP_CD and kinase subfamilies.

The receptor diversity discovered in this work is complemented by diversity in the kinases with which they interact, as elucidated by Kristin Wuichet in the Zhulin lab. A thorough understanding of chemotaxis function and evolution relies on characterizing the receptor / kinase interaction. To that end, an algorithm was developed to associate receptor / kinase pairs, which facilitated reconstruction of the evolutionary history of chemotaxis almost to its origin when a sensor kinase proteins split into separate parts.

Research advance 3: Development and Implementation of Cheops, a database of chemotaxis pathways.

The Cheops (Chemotaxis operons) database rests on the foundation of the Microbial Signal Transduction (MiST) database built by Luke Ulrich in the Zhulin lab. MiST focuses on one- and two-component systems, two of the three major categories of signal transduction modules found in prokaryotes. Cheops completes the work of MiST by adding chemotaxis. Cheops presents the results of the sensor / kinase correlation algorithm and the information about receptor and kinase diversity in an integrated and intuitive way that is useful both for experimentalists interested in guiding their research in a particular model system and for computational biologists interested in gaining a better understanding of the function and evolution of the chemotaxis signaling module.

CHAPTER 1

INTRODUCTION

Since the discovery that DNA is the molecule that carries genetic information [1-3], there have been fifty years of progress in the reductionist scientific program, studying individual genes and their protein products to determine their function and how they contribute to the living cell. Since the mid-1990s, whole genome sequences have been determined at an accelerating rate. There are now hundreds of sequenced prokaryotic genomes and dozens of sequenced eukaryotic genomes. High-throughput functional genomics techniques are now generating experimental data about large numbers of genes and proteins simultaneously, and new techniques are being developed every day. The challenge to biology now is to integrate legacy single-gene experiments from model organisms with comparative and functional genomic data to provide a complete understanding of how living systems function and evolve. That task is easier than it might have been [4]. The field of systems biology was born out of the realization that biological systems can be broken down into functional modules [5] because natural selection favors modular networks that can rewire themselves in response to changing environments [6]. Strongly connected non-modular networks are brittle in the face of such change and do not survive this evolutionary challenge.

The focus of this research project is the functional module that enables prokaryotes to navigate their environment in response to attractant and repellent chemical stimuli. This behavior, called chemotaxis, was first noticed in the late nineteenth century by microscopists who observed bacteria accumulating around air bubbles trapped underneath microscope slides [7]. During the explosion of molecular biology in the 1960s, the molecular basis of this ability was determined by genetic experiments in the

bacterium *Escherichia coli* by pioneers like Julius Adler [7]. The reason for the interest of those pioneering molecular biologists in chemotaxis was their hope that the molecular basis for this behavior in prokaryotes might be the same as that of more complicated behaviors exhibited by humans and other multi-celled organisms. Their hopes were unfounded at the molecular level, since the molecules involved in chemotaxis are almost exclusively restricted to prokaryotes. Chemotaxis has nevertheless become an important model system for understanding signal transduction at the molecular level. With the birth of systems biology, it is becoming clear that understanding chemotaxis as a functional module will in fact guide our research into more complicated systems.

1.1 Types of Motility found in Prokaryotes

The two best-studied chemotactic organisms, *E. coli* and *Bacillus subtilis*, both exhibit flagellar motility in a liquid environment [8]. More specifically, they have peritrichous flagella, meaning they have 4 to 8 flagella distributed around their cell body that form a bundle that propels the cell forward when they all spin in the same direction, counter-clockwise (CCW) as seen from behind. When at least one of the flagella changes direction to spin clockwise (CW), the bundle flies apart, causing the cell to tumble and randomly reorient. These tumbling events are of short duration [9]; all the motors quickly return to spinning CCW, the bundle reforms, and the cell moves off in a new direction. In a uniform environment, this alternation of runs and tumbles leads the cell to perform a random walk. In a gradient, however, the cell controls the duration of its runs, lengthening them when traveling towards a positive stimulus (up an attractant gradient or down a repellent gradient), and shortening them when traveling towards a negative stimulus (down an attractant gradient or up a repellent gradient). The result is a biased random walk with net movement towards a more optimal environment.

Chemotaxis can control cellular motility systems of a variety of types different from the dominant paradigm just described [10]. The peritrichous flagella of *Sinorhizobium meliloti* rotate only in one direction, in a CW bundle, and the cell changes direction when individual flagella cause the bundle to fly apart by changing speed [11,12]. *Pseudomonas aeruginosa* has a single polar flagellum that rotates CCW when running forward; to change direction, the sense of rotation is reversed, pulling the cell backward slightly and reorienting it [13]. *Rhodobacter sphaeroides* has a single lateral flagellum that rotates in one direction; to change the cell's direction, the flagellum stops rotating, forming a tight coil near the cell body which then rotates slowly [12]. In spirochetes, one or more flagella are located at both ends of the cell and sheathed inside the periplasm by the outer membrane; the cell corkscrews through viscous media by rotating the flagella at each pole in opposite directions, changing the direction of swimming by reversing direction of both flagella [14].

In addition to flagellar motility through liquids, *P. aeruginosa* can also move over solid surfaces using a mechanism called twitching motility which is generated not by flagellar rotation but by the extension and retraction of Type-IV pili [15]. Separate chemotaxis systems control flagellar and twitching motility in *P. aeruginosa*. *E. coli* and many other organisms have the ability to swarm over surfaces [16]. The mechanism of swarming motility involves the generation and collective motion of hyperflagellated cells [17]. It requires the chemotaxis system to function, but how or whether the direction of swarming is controlled by chemotaxis is poorly understood [18].

Chemotaxis modules have been found in some cases to control cellular systems unrelated to motility. *Rhodobacter centenum* has two such modules, one that controls flagellar biosynthesis [19], and another that regulates the development of starvation-resistant cysts [20]. One of the eight chemotaxis modules in *Myxococcus xanthus* has been found to control expression of developmental genes [21]. The response regulator of

a third chemotaxis module in *P. aeruginosa* controls biofilm formation via its diguanylate cyclase activity [22].

What enables chemotaxis to control such diverse outputs is its modularity. The output of the chemotaxis network is the level of phosphorylation of one protein, the response regulator CheY. The different systems controlled by chemotaxis respond mainly to the concentration of phosphorylated CheY (CheY~P) and are not strongly wired to the rest of the chemotaxis functional module. An issue that confounds or at least complicates this apparent simplicity is the possibility of crosstalk between modules in organisms with multiple chemotaxis pathways. A central goal of this research is to trace the evolutionary history of chemotaxis with systems biology issues like modularity, robustness, and evolvability in mind. A good starting point is to discuss the evolutionary origin of chemotaxis.

1.2 Chemotaxis evolved from a simpler two-component system

Chemotaxis evolved from a simpler two-component system (TCS) [23-25]. For a long time two-component systems have been the paradigm for signaling in prokaryotes [26]. The two proteins in a TCS are a sensor kinase and a response regulator (Figure 1.1A). The sensor kinase, or Class I histidine kinase, is usually membrane-bound with an extracellular sensory domain. When the sensory domain binds its cognate ligand, it changes conformation, signaling to the kinase domain in the cytoplasm to change the rate of phosphorylation of a conserved histidine residue. The phosphate group is quickly transferred from that histidine to a conserved aspartate residue in the receiver domain of the output protein. The receiver domain changes conformation in response to the negative charge of the phosphate group, and that change is sensed by an output domain which performs some function. In 75% of TCS, the output protein is a transcription factor, i.e. the receiver domain is coupled to a DNA-binding domain, so the TCS as a whole affects gene regulation [24].

The Class II histidine kinase involved in chemotaxis, called CheA, originated when a Class I sensor kinase split into two separate proteins (Figure 1.1B). The split also involved the birth of the CheW scaffolding protein which mediates the interaction between the kinase and the sensor. The kinase actually has a domain homologous to the CheW scaffold domain that participates in the scaffolding interaction [27]. So the breakup of the sensor kinase into separate parts was accompanied by the birth and duplication of a scaffold protein to mediate the interaction. The order of these evolutionary events is unknown, although it is probable that the birth of CheW occurred first. Otherwise there would be no mechanism to mediate the interaction of the separated sensor and kinase domains, and the system would have lost its functionality.

The birth of separate sensor and kinase proteins and mediation of their interaction by a new scaffold protein had a clear selective advantage; it made possible one of the primary functions of chemotaxis, namely the ability to integrate disparate signals from multiple sensory inputs [28-30]. Figure 1.2 shows a recent image of an array of sensory receptors embedded in the membrane of *E. coli* associated with an array of scaffold and kinase proteins on the cytoplasmic side of the membrane. These proteins form a large array at the cell pole in *E. coli* [31], and similar arrays form in the same location in many different species of bacteria [32]. This structure is central to the mechanism of signal integration. Suppose that one receptor senses a strong chemoattractant, another receptor senses a weak chemoattractant, and a third receptor senses a chemorepellant. In an environment containing all three chemicals, the cell has to choose one direction of travel based on integrating those multiple sensory inputs. Experimental work in *E. coli* has focused on understanding details of this signal integration mechanism using the serine and aspartate receptors, Tsr and Tar [33-35]. While their pairwise interactions have been fairly well characterized, how signal integration functions at the level of the entire chemoreceptor array is an important and still poorly understood question that has been a major focus of this research.

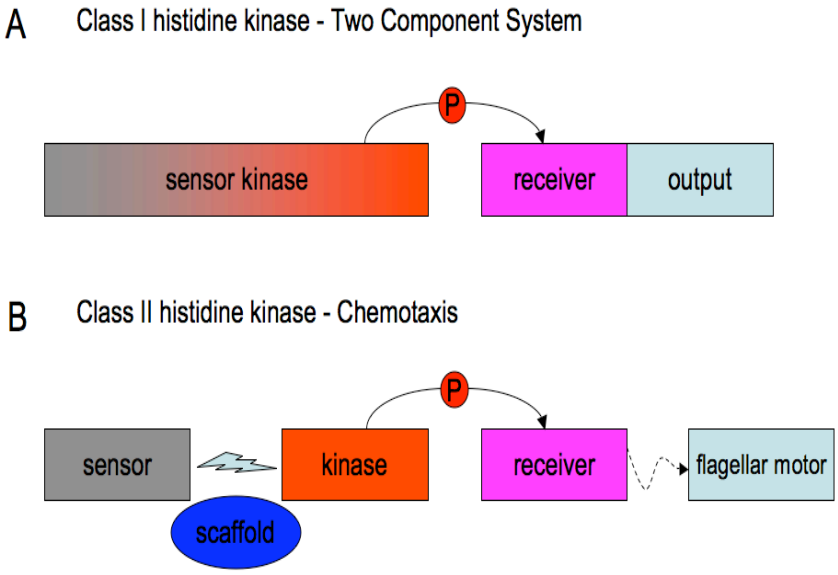

Figure 1.1 Chemotaxis evolved from a two-component system (TCS). (A) The classic TCS consists of a sensor histidine kinase that transfers a phosphate group to the receiver domain of an output protein. The phosphate group induces a conformational change that signals the output domain to perform its function, usually gene regulation by DNA-binding. (B) The Class II histidine kinase that functions in chemotaxis originated when the Class I sensor kinase split into separate sensor and kinase proteins, MCP and CheA, respectively. Their interaction is mediated by a scaffold protein, CheW. In chemotaxis, the receiver domain is a stand-alone protein, CheY, that diffuses through the cell to interact with its output. In *E. coli*, that output is the flagellar motor.

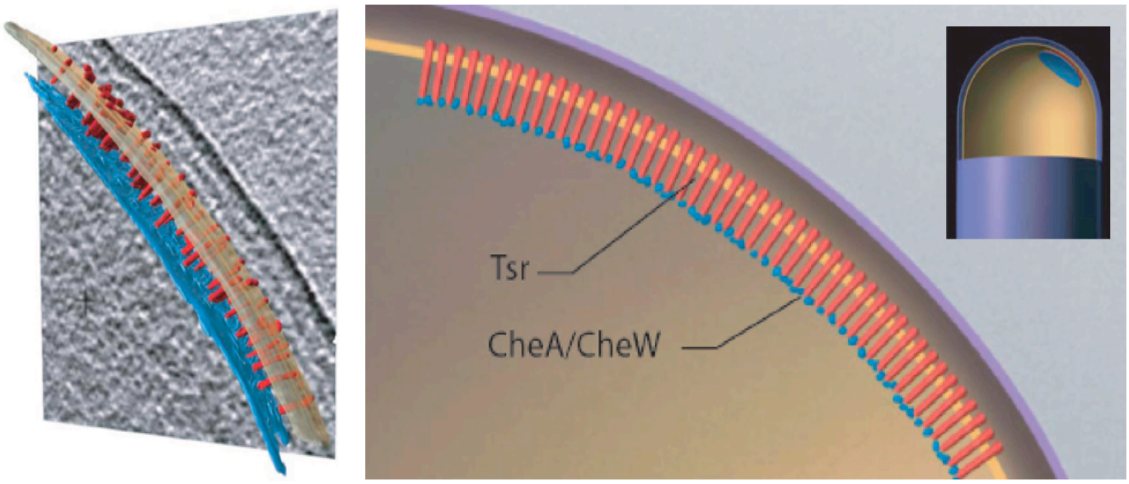

Figure 1.2 The sensor, scaffold, and kinase proteins co-localize to the polar region of the membrane in many species [31,32]. Seen here is a recent cryo-electron microscopy image of the chemoreceptor array in *E. coli* [36]. (left) Cryo-EM image with membrane (tan), sensors (red), and kinase and scaffold (blue) density sketched in for clarity. (right) Schematic representation of the chemoreceptor array. Image adapted with permission from [36]. Copyright National Academy of Sciences, 2007.

1.3 Central Role of the Sensor and Kinase Proteins

The interaction between the sensor and the kinase is central to chemotaxis signal transduction. A comparative genomic analysis of chemotaxis should therefore focus on those two proteins. The sensor proteins are called methyl-accepting chemotaxis proteins (MCP) for reasons that will soon be made clear. My research focus has been analysis of the conserved cytoplasmic signaling domain of MCPs [37]. My colleague, Kristin Wuichet, has focused on analysis of the Class II histidine kinase, CheA [25]. In this thesis, I will present the details of my analysis of the MCP cytoplasmic domain and an overview of Kristin's kinase analysis, and then explain how integration of the comparative genomic analysis of sensor and kinase allows deep insights into the function and evolutionary history of chemotaxis.

1.4 Chemotaxis in *E. coli*

Bacteria are too small to sense a spatial gradient of chemicals of interest; instead they sense a temporal gradient [38]. They retain a memory of the concentrations of interesting chemicals from a few seconds ago, compare it to their current concentration, and decide whether to continue swimming or to tumble and change direction. What is the molecular basis of this bacterial memory circuit? There are two orthogonal routes of signal transduction in the chemotaxis network of *E. coli* (Figure 1.3). First is the excitation pathway. Information about ligand binding at the MCP sensory domain controls the direction of rotation of the flagellar motor by traveling through the HAMP linker domain, the MCP cytoplasmic domain, the CheW scaffold, the kinase CheA, the receiver protein CheY, and the phosphatase CheZ. Ligand binding at the sensor modulates the rate of kinase activity, and the flagellar motor senses the level of phosphorylation of CheY after it has left the chemoreceptor array and diffused through the cytoplasm to the location of the motor. CheZ, which acts as a CheY~P phosphatase [39,40], localizes to the chemoreceptor array via a modified form of CheA called CheA

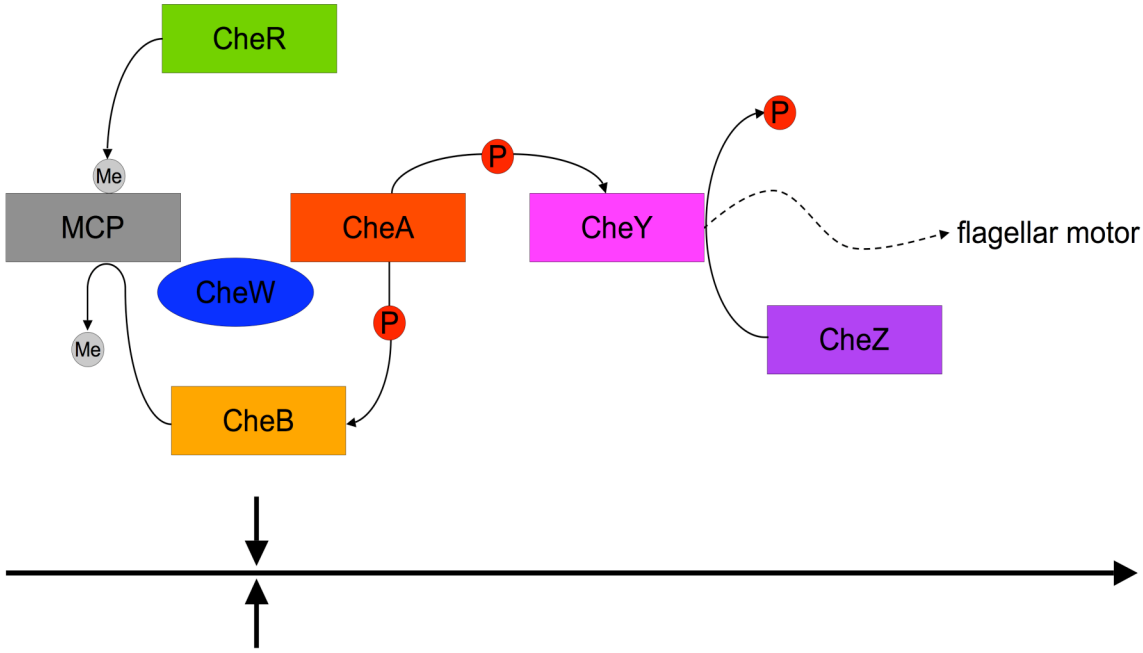

Figure 1.3 The chemotaxis signal transduction network of *E. coli*. Arrows at bottom indicate the two orthogonal routes through the pathway. The long arrow indicates the excitation pathway whereby ligand binding at the MCP sensory domain alters the activity of the kinase CheA, which changes the level of phosphorylation of the receiver protein CheY. The flagellar motor responds to the phosphorylation state of CheY and thus to the ligand binding event. The short arrows indicate the adaptation pathway whereby CheR and CheB alter the methylation state of the MCP cytoplasmic domain.

short (CheAs) [39]. The effect of CheZ localization is probably to establish a uniform concentration of CheY~P throughout the cell downstream of the array so that all the motors sense the output of the chemotaxis system to the same degree [41,42]. CheZ also keeps the cell responding to current stimuli by timely turnover of CheY~P.

Second is the adaptation pathway consisting of the methyltransferase CheR [43-46] and the methylesterase CheB [47-49]. Methyl-accepting chemotaxis proteins are so named because CheR adds and CheB removes methyl groups from specific glutamate residues in the MCP cytoplasmic domain [50]. Some of these residues are encoded as glutamines in the MCP gene; in *E. coli*, CheB has a deamidase activity that post-translationally modifies them to glutamate to activate them for the methylation /

demethylation cycle [51]. The effect of methylation on the conformational dynamics of the MCP is the molecular foundation of the memory storage mechanism. The methylation state of the MCP affects the kinase activity of CheA. Fine control of the balance between the activities of CheR and CheB generates switch-like behavior in the system, so that it outputs either a very low or very high concentration of CheY~P [52].

1.5 Molecular Basis of Adaptive Feedback

There is an interesting epistemological conflict between what I will call "old school" molecular biology and "new school" systems biology. They have different explanations for the adaptive feedback mechanism at the heart of the chemotaxis system.

1.5.1 Hypothesis from Biochemistry

The "old school" hypothesis from biochemistry is as follows: it has been known for a long time that the demethylation activity of the CheB methylesterase depends on its phosphorylation state. CheB is a two domain protein; it has a methylesterase domain and a receiver domain like CheY. When the CheB receiver domain is phosphorylated, CheB removes methyl groups much more rapidly from glutamate residues in the MCP cytoplasmic domain than when it is not [48,49,53].

In *E. coli*, when a chemo-attractant binds to the MCP sensory domain, the MCP shifts into a conformational state that results in a reduced level of CheA activity [50], thereby decreasing the level of both CheY~P and CheB~P in the cell. When the flagellar motor senses that the level of CheY~P is low, it keeps turning counter-clockwise, continuing to run forward [54]. But there is a negative feedback loop: because there is less CheB~P in the cell, demethylation of the MCPs slows down. CheR is constitutively active, so when CheB activity is reduced, the level of methylation of the MCPs increases. In *E. coli*, adding methyl groups to MCPs increases the level of kinase activity [50]. In summary, adding attractant decreases kinase activity, but via the CheB feedback loop,

after a delay, kinase activity returns to its previous level because of increased methylation. The cell is then primed to respond to even higher concentrations of chemo-attractant and thus to continue swimming up the gradient.

There are several experimental ways to measure kinase activity. Two classic methods are video tracking of swimming cells [9] or of tethered cells [38,55]. In the tethered cell assay, cells are attached to a microscope slide by flagella that have been sheared by sonication. Some fraction of the tethered cells are able to rotate in both directions. By tracking software or tedious manual analysis, the ratio of time spent in the CW vs CCW rotational state (CW bias) can be measured. The data generated by swimming cells is tumbling frequency, the number of tumbling events that occur in a given period of time. CW bias and tumbling frequency are equivalent measures since tumbling events have an approximately constant duration of 0.14-0.2 events/sec [9,56]. A more recent method that produces better quantitative data is to calculate the level of CheY~P *in vivo* indirectly by measuring fluorescence resonant energy transfer (FRET) between CheY and its phosphatase CheZ [57].

Figure 1.4 shows a schematic time course of how CheA activity in a population of *E. coli* cells responds to the addition of attractant. There is a steady state of kinase activity before attractant addition. Afterwards, there is an immediate steep response; the level of kinase activity drops rapidly. Over time, the population slowly adapts back to its pre-stimulus steady state even though the attractant concentration remains high. Recalling that the function of chemotaxis is to control the duration of each run, the curve has two important characteristics. First, there is high gain: the amount of change in kinase activity is large in comparison to the amount of chemo-attractant bound at the receptors. Second, the adaptation time, encoded in the feedback loops, binding constants, and enzymatic rates within the network of interacting chemotaxis proteins, is important. Both gain and adaptation time play a role in determining the length of runs, or equivalently the time between tumbling events.

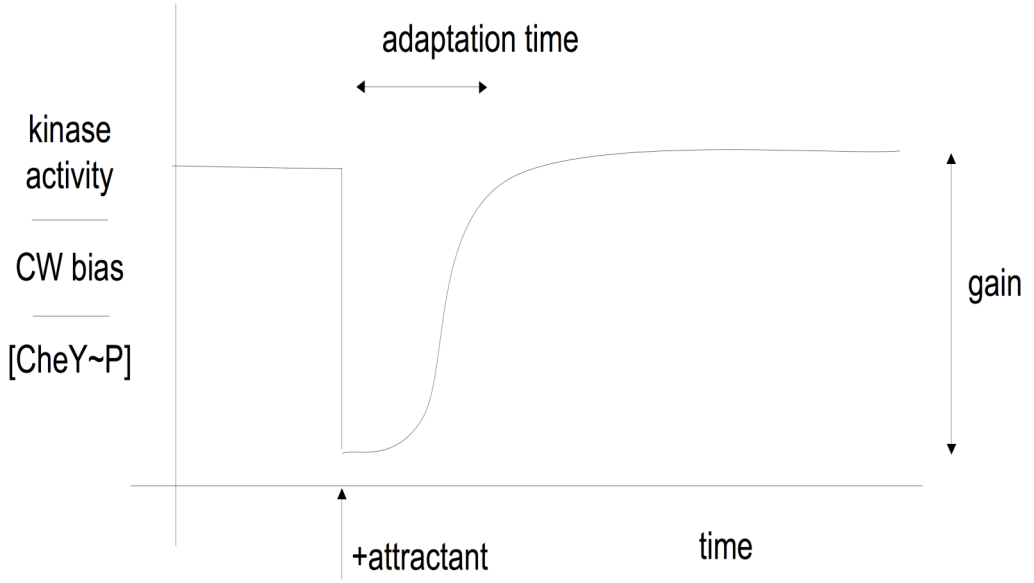

Figure 1.4 Kinetics of adaptation in *E. coli*. In an environment devoid of chemostimuli, a population of cells exhibits a steady-state level of kinase activity, which can be measured in several ways (see text). When attractant is added, *E. coli* cells rapidly decrease the level of kinase activity. Over time, the kinase activity returns to its pre-stimulus steady-state. The duration of a run is related to both the initial gain and the adaptation time.

1.5.2 Hypothesis from Modeling

The "new school" hypothesis about the molecular basis of adaptive feedback is as follows. In 1997, Barkai and Leibler performed a robustness analysis of the chemotaxis network of *E. coli* [58]. Their hypothesis was that only network architectures that are robust against gene expression noise will be maintained by natural selection. An organism containing a system that performs its function only when it has exacty the right ratio of all interacting proteins will survive less frequently than one containing a system that still functions properly even when there is noise in gene expression and variability in protein concentrations.

So Barkai and Leiber created a mathematical model of chemotaxis in *E. coli* looking for network architectures that exhibited robustness. They concluded that the only way for chemotaxis to be robust is for demethylation of the MCP by CheB to occur only when the MCP is in a kinase-activating conformational state. They called this

mechanism "activity-dependent kinetics." In 1999, with Alon and Surette, they tested their hypothesis experimentally by engineering cells that expressed a CheB lacking a receiver domain [59]. They showed that such cells could still adapt precisely, so that the feedback loop by phosphorylation of CheB was not necessary for precise adaptation. They then assumed that activity-dependent kinetics was the reason for precise adaptation by this constitutively active CheB. Actually, in later models, it has been assumed not only that CheB acts only on an active conformation of the MCP but also that CheR acts only on an inactive conformation of the MCP [60-67].

What then is the role of CheB phosphorylation? Later modelers have argued that the modulation of CheB activity by CheA plays some role in controlling gene expression noise [67]. The ten-year-old prediction about the mechanism of interaction between CheB and the MCP cytoplasmic domain has still not been confirmed structurally. Because of the complex environment of the chemoreceptor array, nobody has yet characterized the structural interaction between CheB and the MCP or between CheR and the MCP. If the prediction of Barkai and Leibler is someday confirmed, it will be a remarkable achievement that a network model based on ideas about robustness from a systems perspective could make such a precise prediction about the structural interaction between two proteins. Proof of their conjecture would show the power of systems biology. The current uncertainty in the field about the molecular basis of the adaptive feedback mechanism in chemotaxis highlights the tension between the "new school" explanations that will increasingly come from systems biology and modeling and the "old school" explanations that come from molecular biology and biochemistry.

1.6 Population Variability

Steven Chu, the Nobel-prize winning director of Livermore laboratory, has a crude metaphor that illustrates why population variability is an important factor to consider when modeling the behavior of biological systems [68]. He says that the average

human has one testis and one ovary, so to model human sexual dynamics, a good starting point from this average data would be to assume that humans are all hermaphrodites who can self-fertilize. The model is hilariously wrong because of the presumption that the average actually exists within one individual, when in fact there is diversity within the population and no individual embodies the average. Since population averages are problematic, we should consider population variability and diversity in our models and experiments. I will present here two cases where population variability has had an important impact on our understanding of chemotaxis. The implications of the first case are simple and straight forward, while those of the second are not yet resolved.

As described above, one of the remarkable properties of chemotaxis is its ability to amplify a small change in ligand concentration into a large change in kinase activity. An important task is to pinpoint where in the chemotaxis network architecture this gain originates. The cytoplasmic face or 'C-ring' of the flagellar motor has ~34 subunits [54,69]. Part of the gain in the system is generated by allosteric interactions between these subunits upon binding of CheY~P. Measurements based on population averages implied there was little cooperativity at the flagellar motor, yielding a Hill coefficient of ~2.5 [56]. Single-cell measurements, on the other hand, yielded a Hill coefficient in individual flagella of 10.3±1.1, and proved that the error in the earlier experiments was caused by population averaging [70]. Thus a significant portion of the gain in the chemotaxis network architecture occurs at the flagellar motor, and this fact was obscured for some time by the smoothing effect of population-averaged measurements.

The robustness result from Barkai and Leibler [58] was based on a deterministic simulation, and its experimental confirmation was based on measurements of tumbling frequency averaged over a population of 100-400 cells [59]. Figure 1.5A, reprinted from [59], shows the steady state chemotactic behavior of a cell population and a time course of adaptation after attractant addition (Compare to the schematic time course in Figure 1.4). Figure 1.5B, adapted from [64], shows that a single cell exposed to a uniform pre-

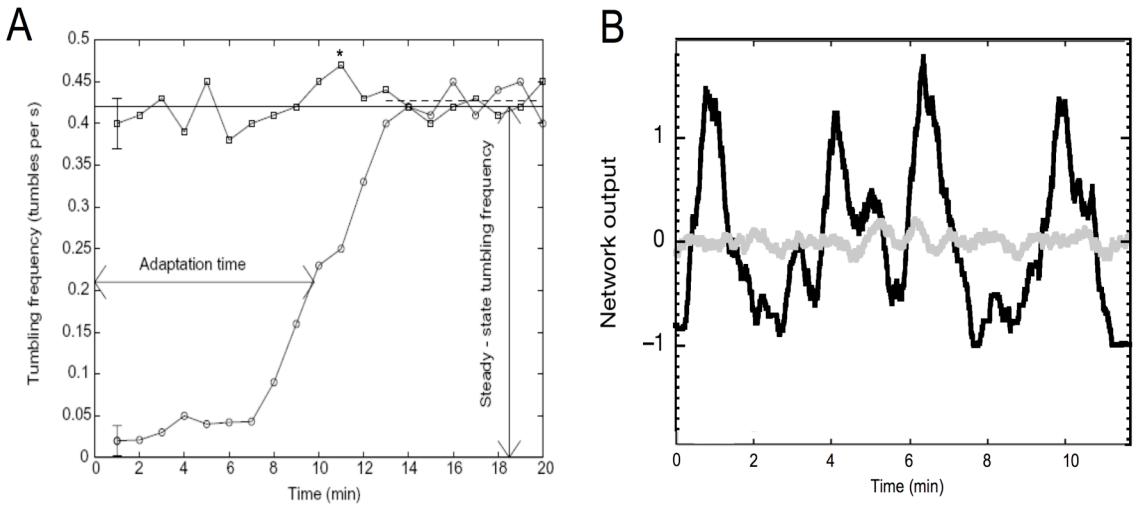

Figure 1.5. Comparison of population-averaged with single-cell measurements of kinase activity in *E. coli*. (A) Here kinase activity is measured in terms of tumbling frequency of whole cells. Each data point represents the average behavior of 100-400 cells tracked for 10 seconds by video microscopy. Squares: unstimulated cells in chemotaxis buffer. Circles: cells stimulated with saturating attractant (1mM L-Asp). (*) For comparison to (B), this data point has a network output of ~0.13. Reprinted by permission from Macmillan Publishers Ltd: Nature [59], copyright 1999. B) Here kinase activity is measured in unstimulated cells by "network output," i.e. (CW bias - <CW bias>) / <CW bias>. Black line; wild-type cell with mean CW bias of 0.2 and tumbling frequency of 0.4 s^{-1}. Grey line: mutant cell with 10x WT [CheR]. Adapted by permission from Macmillan Publishers Ltd: Nature [64], copyright 2004.

stimulus environment never settles into the steady state of kinase activity observed at the population level in Figure 1.5A. Instead there are large variations in CW bias (Network output in this figure is a normalized measure of CW bias, specifically (CW bias - <CW bias>) / <CW bias>). Interestingly, the grey line shows that over-expressing CheR level to 4 times its wild-type level damps out the variability and allows even single cells to reach a steady state. The lack of a steady-state network output means that individual cells vary their run length. In an environment lacking chemoattractants, this search strategy may explore the space better than would the CheR mutant where each cell has a constant run length [64].

The implications of this analysis have not yet been fully appreciated by the chemotaxis modeling community. Stated most forcefully, the single cell measurements call into question the robustness analysis based on population average measurements and deterministic modeling, since [CheR] is in some sense maintained at a fine-tuned level by selection for the optimal search strategy in an unstimulated environment. While this strong formulation of the problem may be an overstatement, the issue clearly warrants further study.

1.7 Diversity in Chemotaxis Networks

1.7.1 Chemotaxis in *Bacillus subtilis*

Figure 1.6 shows the architecture of the chemotaxis system in *B. subtilis*. It contains all of the components found in *E. coli* except that the CheZ phosphatase is replaced by the CheC phosphatase, which has a different evolutionary origin and molecular mechanism of action [55,71-74]. *B. subtilis* has a second scaffold protein, CheV, in addition to CheW, that is essential for chemotaxis [75]. CheV also has a CheY-like receiver domain, so there may be active and inactive states of the CheV scaffold depending on its phosphorylation state. *B. subtilis* also has the protein CheD which acts as a receptor deamidase [76]. Some of the glutamate residues in the MCP cytoplasmic domain that are active in the adaptation mechanism are initially encoded as glutamines [77,78]. In *E. coli*, CheB demidates these residues from glutamine to glutamate to activate their role in adaptation [79], but CheD plays that role in *B. subtilis*.

All of the additional proteins in the *B. subtilis* chemotaxis network participate in feedback loops not present in the *E. coli* network. The feedback loop via phosphorylation of CheV is apparent, but it is also the case that interaction between CheC and CheD activates the CheY~P phosphatase activity of CheC [80,81]. This interaction represents a possible feedback loop from CheY through CheC and CheD to the chemoreceptor array.

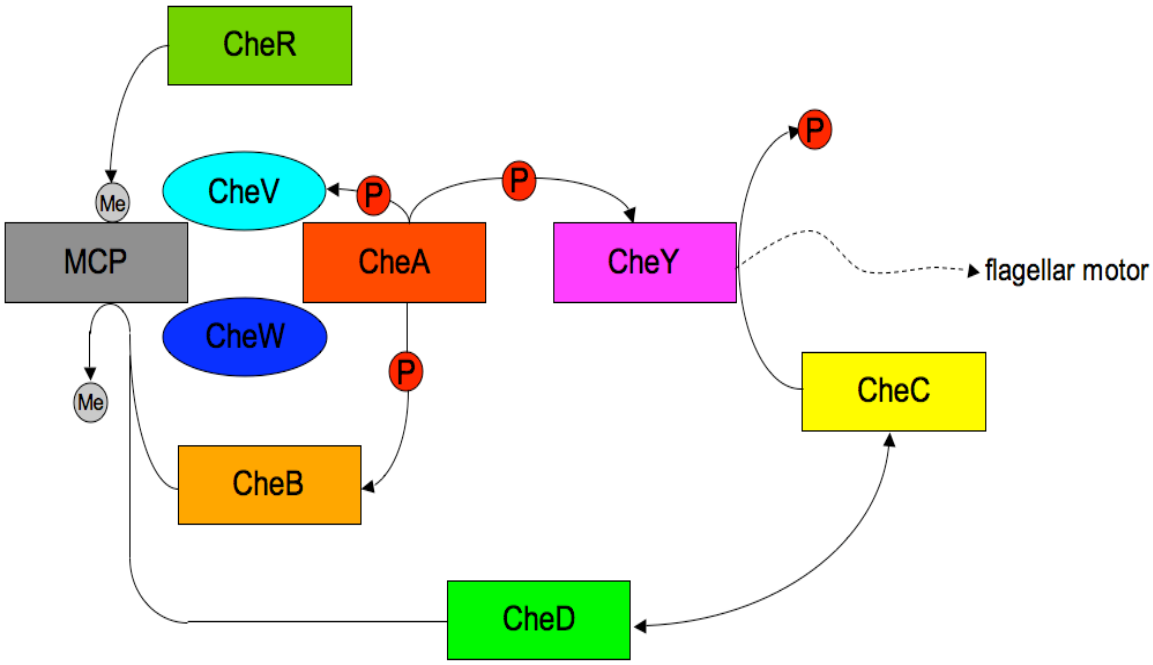

Figure 1.6 Network architecture of chemotaxis in *Bacillus subtilis*.

1.7.2 Evolvability and Robustness

The question of why there is extra feedback in the network architecture of *B. subtilis* chemotaxis compared to *E. coli* is important. A recent simulation exploring the evolution of complexity in a generic three protein signaling network provides a possible explanation [82]. The simulation showed that subjecting a simple signaling pathway to the mechanisms of genome evolution, including gene and genome duplication, led to an increase in the minimum number of proteins in the network. Unless there was strong selection pressure against increasing the number of proteins in the genome, the original minimally robust network ended up gaining proteins and having extra feedback loops, generating a "reservoir of robustness" in the network. If an organism with such an especially robust network later faced a period of strong selection pressure, extraneous feedback loops might be removed and still leave a network architecture robust enough to maintain adequate function. Perhaps the chemotaxis network in *E. coli* lacks such extra feedback loops because of its binge-and-purge lifestyle. In the rich environment of the

gut, *E. coli* grows as fast as possible, which generates selection pressure to remove extraneous proteins that slow the growth rate. Thus its chemotaxis network has been streamlined and is minimally robust [67]. In chapter 4 we will return to this notion that network architectures with more feedback than necessary to perform their function represent a reservoir of robustness that imparts the organism with evolvability, or the ability to remain functional even under strong selection pressure to reduce the size of the genome.

1.8 Methyl-accepting Chemotaxis Proteins

The main focus of this research project is a comparative genomic analysis of the sensory receptors at the beginning of the chemotaxis signal transduction cascade. From our overview of the adaptation mechanism in *E. coli* chemotaxis it should be clear why analyzing the cytoplasmic domain of the receptors is important. The MCP cytoplasmic domain interacts with the scaffold, the kinase, and all the adaptation enzymes. The methylation sites central to the adaptation mechanism are there. To prepare for our analysis it is important to set forth what is already known about receptor structure and function.

1.8.1 Domain Organization and Membrane Topology

MCP sequences typically consist of a sensory domain, a HAMP linker domain, and a signaling domain that interacts with the scaffold CheW and kinase CheA. The HAMP and signaling domains are always cytoplasmic, but the membrane topology of the sensory domain varies. Figure 1.7 shows a classification of MCP membrane topology into four major classes [83]. Sensory class I MCPs have a periplasmic sensory domain anchored by an N-terminal transmembrane (TM) helix and connected by an internal TM helix to the HAMP linker and signaling domains. Most MCPs, including the Tar, Tsr, Trg, and Tap receptors of *E. coli*, have this sensory topology [84]. Sensory class II MCPs

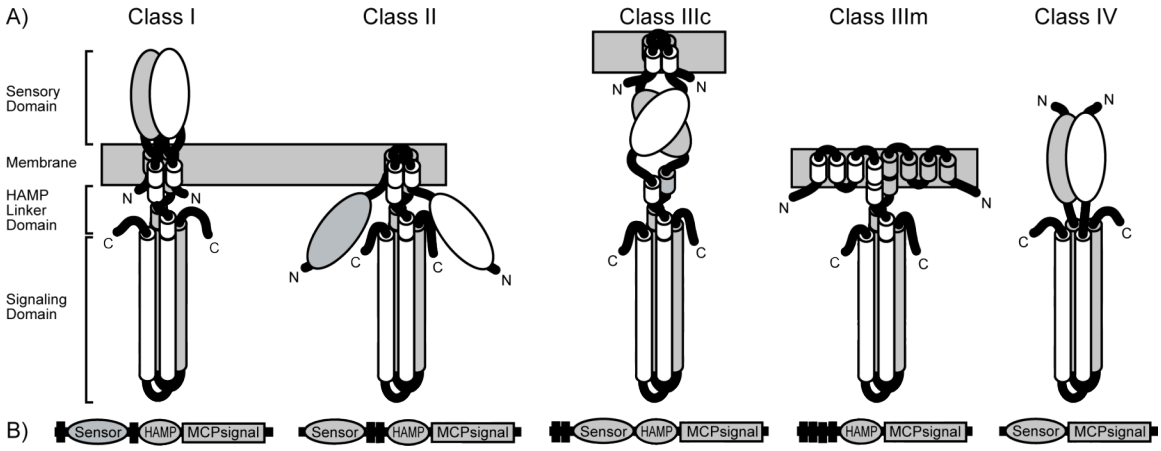

Figure 1.7 Differing membrane topology divides MCPs into four main sensory classes. (A) Schematic representation of the 3D structure of MCP dimers of different sensor classes. Oval domains are sensory domains of varied secondary structure. Cylinders represent alpha-helical and coiled coil regions. MCP monomers are differentiated by grey and white coloring. (B) MCP sensor class can be determined from domain architecture where transmembrane regions and domains are well-predicted. Transmembrane regions are indicated by black boxes. Periplasmic or extracellular domains are white; cytoplasmic domains are grey.

have an N-terminal cytoplasmic sensory domain connected by an internal TM helix to the HAMP linker and signaling domains. The Aer aerotaxis receptor of *E. coli* is an example of a Class II sensor [85]. Since an earlier classification of MCP sensor classes [83], many more MCP sequences have become available, and we have split sensor class III into two subgroups. Sensory class IIIc MCPs are anchored at their N-terminus by a TM helix, downstream of which are a cytoplasmic sensory domain and the HAMP linker domain and cytoplasmic signaling domain. Sensory class IIIm MCPs are like class IIIc MCPs except that the sensory domain is membrane-bound rather than cytoplasmic. The Htr8 aerotaxis receptor of *Halobacter salinarum* is an example of a sensory class IIIm receptor [86]. Some MCPs are hybrids of class II and class III, containing a periplasmic sensory domain separated by a TM helix from an additional cytoplasmic sensory domain [87]. Sensory class IV MCPs are entirely cytoplasmic; they lack TM helices and usually also

HAMP domains. The oxygen sensor HemAT from *B. subtilis* is an example of a Class IV sensor [88].

1.8.2 The Cytoplasmic Domain

The MCP cytoplasmic domain is an anti-parallel four-helix coiled coil [89,90]. A coiled coil consists of alpha helices supercoiled around each other so that two turns of a supercoiled alpha helix is exactly seven residues [91]. Each group of seven residues in a coiled coil is termed a "heptad" and its residues are labeled a-b-c-d-e-f-g. The a and d residues tend to be hydrophobic and are the knobs that give the coiled coil its stability. Coiled coil structures with two, three, four, and even five helices are known [92,93]. Shown in Figure 1.8A are two known structures of the MCP cytoplasmic domain. On the left is Tsr, the serine receptor from *E. coli*, crystallized in 1999 [90], and on the right is a receptor from *Thermotoga maritima*, TM1143, crystallized more recently [94]. Both receptors show the anti-parallel four-helical bundle structure of the domain. Figure 1.8B shows a schematic diagram of how the heptad register maps onto the four-helical bundle formed by the MCP dimer. At the center of the bundle is an a-d knob layer, while the b, c, and f heptad registers are on the surface of the bundle. Figure 1.8C shows a representative knob in Tsr. A knob residue on one helix projects into a pocket of four hole residues on an adjacent helix. Hole residues are themselves often knobs with respect to an adjacent helix. Knob residues stabilize helix interactions in coiled coil proteins. A major finding of this research project was a characteristic arrangement of knob layers in the MCP cytoplasmic domain that has implications for the MCP signaling mechanism (see section 3.2).

Before this research, the MCP cytoplasmic domain was divided into two functional regions with unclear boundaries. At the base is the signaling subdomain, highly conserved [89] because it interacts with the scaffold CheW, the kinase CheA, and other MCP dimers in the chemoreceptor array. Residues in the signaling subdomain can

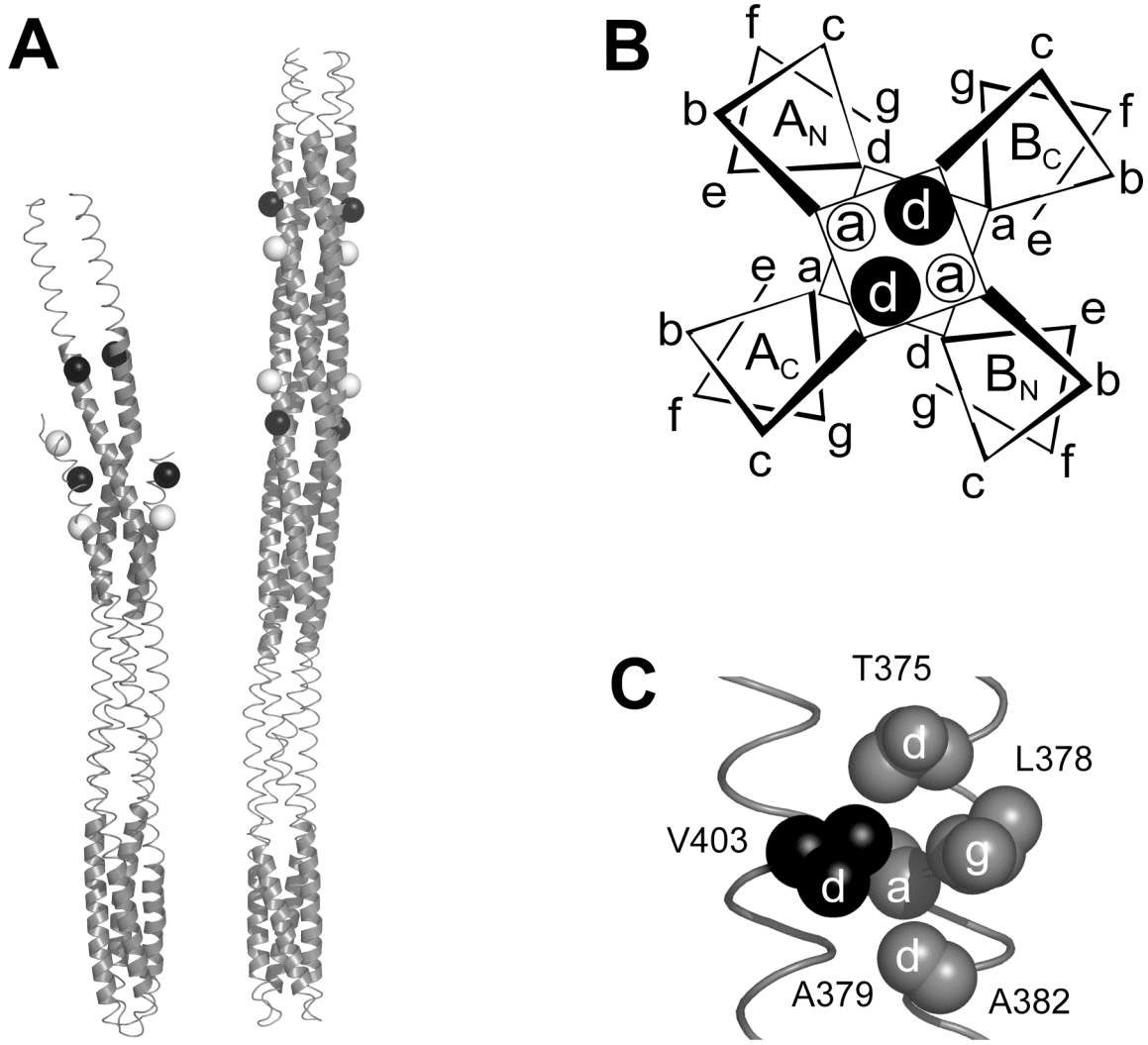

Figure 1.8 Structure of the MCP cytoplasmic domain. (A) Structures of the Tsr (left) and TM1143 (right) cytoplasmic domain show coiled coil regions determined by the SOCKET algorithm with a 7.8 Å cutoff [92,93]. Thin ribbons indicate regions where coiled coils were not detected. Experimentally determined sites of methylation are indicated on the structures as spheres, colored black or white if encoded in the gene as glutamate or glutamine, respectively. (B) Schematic representation of the arrangement of coiled coil heptads in a four-helical bundle dimer of the MCP cytoplasmic domain, viewed axially from the top. Monomers A and B each have N-terminal (A_N, B_N) and C-terminal (A_C, B_C) helices joined by a hairpin loop at the base. Heptad registers a and d form a square-shaped knob layer at the core of the bundle. (C) Representative knob (black) into hole (grey) packing in the *E. coli* Tsr protein. Residue numbers and heptad registers are shown.

be partitioned into two classes based on heptad register. The b, c, and f registers on the surface of the bundle are inter-dimer contact sites which mediate higher-order interactions between dimers in the array, while registers a, d, e, and g are intra-dimer contact sites that stabilize the dimer core (Figure 1.8B). In the crystal structure, the Tsr receptor from *E. coli* formed a trimer of dimers based on inter-dimer interactions in the signaling subdomain [90]. The importance of trimer contact sites in mediating receptor interactions has been shown experimentally in *E. coli* [34], but the trimer of dimers structure was not found when TM1143 was crystallized [94]. Instead, TM1143 formed a "hedgerow of dimers" with only pairwise dimer interactions.

At the top of the MCP cytoplasmic domain are the methylation helices where adaptation enzymes act. Experimentally determined sites of methylation in Tsr and TM1143 are indicated in Figure 1.8A as spheres. Black spheres indicate methylation sites encoded in the gene as glutamate. White spheres indicate sites encoded in the gene as glutamine, which must first be post-translationally modified to glutamate by CheB or CheD before they can be methylated.

The MCP cytoplasmic domain is the principal object of study in this research project. I will refer to it in several ways throughout this thesis depending on context. I will often use the abbreviation MCP_CD when focusing on an alignment of the domain. To contrast it with the sensory and HAMP linker domains, I will sometimes refer to the entire MCP cytoplasmic domain as the signaling domain. When referring to the region that interacts with the scaffold and kinase, I will always specify that it is the signaling *sub*-domain.

1.8.3 The HAMP Linker Domain

The HAMP domain is a conserved linker domain found in a variety of transmembrane signaling proteins; its name stems from its presence in histidine kinases, adenylate cyclases, MCPs, and phosphatases [95]. Because of its structural flexibility, the role of the HAMP domain in transmembrane signaling has been difficult to pinpoint, despite significant molecular biological efforts to do so [85,96-106]. Cysteine scanning and sequence analysis showed that the domain probably consisted of two helical amphipathic sequences, AS1 and AS2, separated by a flexible linker [97,98], but the structural interaction between and the functional role of these components was unclear until . Recently, however, the structure of a particularly stable single-domain HAMP protein from an archaeal species was determined via NMR spectroscopy [107]. The structure revealed a parallel four-helical bundle with a non-standard packing arrangement (Figure 1.9). This structure led to the proposal that signaling in the HAMP domain consists of the concerted, gear-like rotation by 26° of all four helices from the non-standard to a standard knobs-into holes packing arrangement. In section 1.8.5, we will assess the impact of this new structure on our understanding of the MCP signaling mechanism.

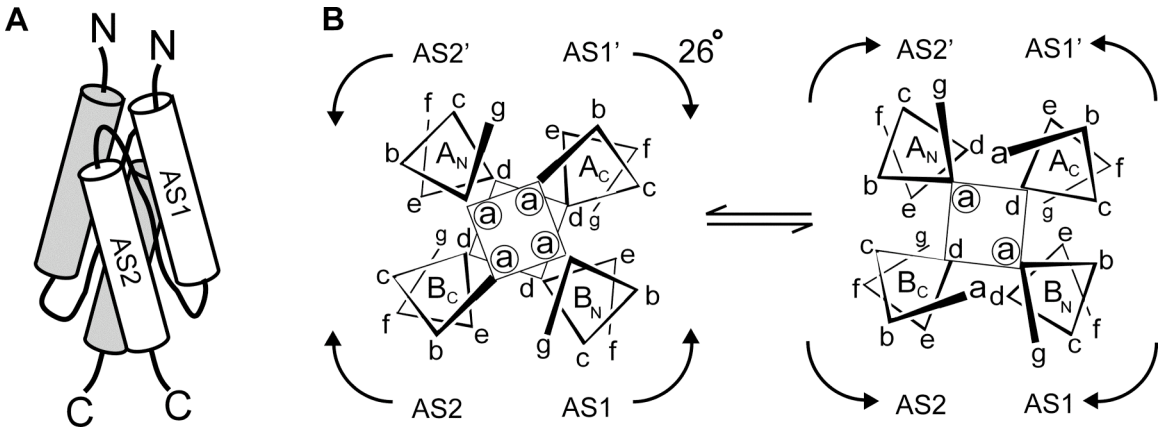

Figure 1.9 Proposed HAMP domain signaling mechanism [107]. (A) The archaeal HAMP dimer forms a parallel four helical bundle coiled coil; each monomer has two helical amphipathic sequences, AS1 and AS2, connected by a disordered flexible linker. (B) At left is the standard knobs-into-holes packing arrangement of heptads in a parallel four helical bundle. Knob layers at the core of the bundle consist of alternating a-a and d-d layers, unlike the mixed a-d knob layers in the antiparallel bundle of the MCP cytoplasmic domain. The packing arrangement actually found in the HAMP structure is shown at right; it is a non-standard arrangement called "complementary x-da" packing, rather than knobs into holes. The proposed signaling mechanism is a concerted, gear-like rotation by 26° of all four helices between the standard and non-standard packing conformations.

1.8.4 Sensory Domains

The MCP cytoplasmic domain is highly conserved because it maintains multiple protein–protein interactions within the chemoreceptor–kinase complex. MCP sensory domains, however, evolve rapidly, are subject to frequent domain birth and death events, and are quite variable in sequence [87]. The lack of good sensory domain models is still an unsolved problem not only in chemotaxis, but in microbial signal transduction in general [24]. About 20% of MCPs have sensory domains identified by Pfam or SMART, while another 20% have the same four-helical bundle structure (TarH / 4HB_MCP) as the four membrane-bound MCPs in *E. coli* [84]. Figure 1.10 shows an array of the sensory domains found in MCPs by MiST. PAS [108] and GAF [109] are ubiquitous sensory domains with a similar protein fold, now called the PAS/GAF fold; PAS and

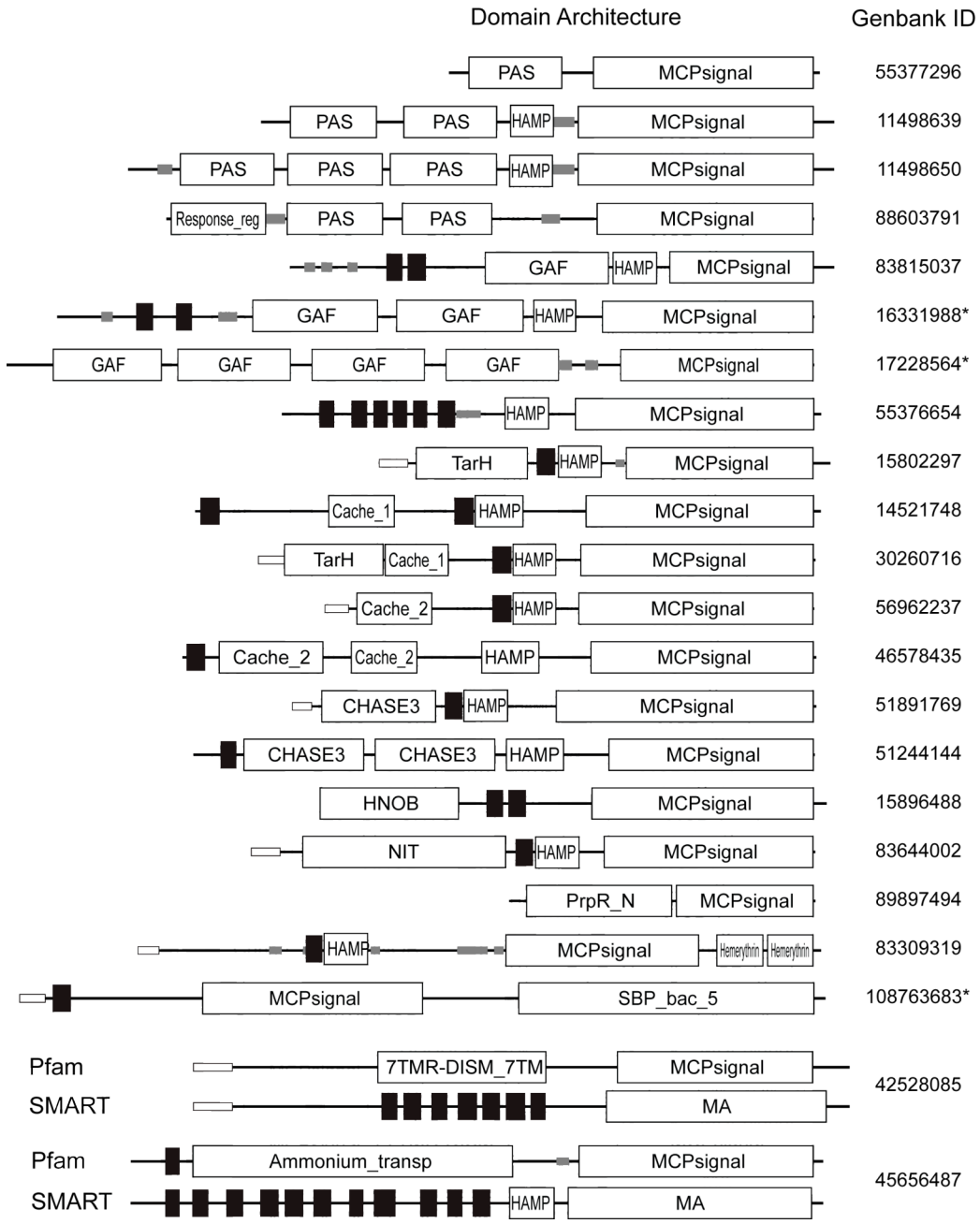

Figure 1.10 Diversity of Sensory Domains in MCPs. All sensory domains are from Pfam domain models, except the GAF domain. The SMART GAF domain model is slightly longer than the Pfam domain model. HAMP domains use the SMART domain model. MCPs containing hemerythrin and SBP_bac_5 sensory domains represent the unusual topology where the MCP signaling domain is N-terminal of the sensory domain. Both Pfam and SMART domain architectures are shown for two MCPs with class IIIm membrane topology. Small grey and white boxes indicate predicted low complexity and signal peptide regions, respectively. Black boxes represent transmembrane regions. Long sequences marked by an asterisk (*) were shortened for display and are not to scale.

GAF are found in both prokaryotic and eukaryotic signaling proteins. Most members of these domain families are cytoplasmic, although a divergent PAS subfamily is exclusively extracellular [110]. In addition to MCPs, where they are always located extracellularly, Cache family domains are found in extracellular subunits of eukaryotic calcium channels that are implicated in signal transduction [111].

A narrow range of signal specificity can be proposed for some sensory domains, for example, the nitrate- or nitrite-responsive NIT domain [112]; however, in most instances, the spectrum of input signals cannot be readily predicted only from analysis of the domain sequence. For example, for most of the MCPs with the TarH fold, nothing is known about their signal specificity, and the diversity of signals sensed by the *E. coli* receptors indicates the versatility of the fold. For the 60% of MCPs in which no known sensory domains are identified by current models, the best that can be done is to predict their sensory topology from the pattern of TM helices (see section 2.12). These receptors contain either known domains not recognized by low-sensitivity models or novel, uncharacterized domains. Further computational and experimental work is necessary to identify and understand the function of novel sensory domains in MCPs.

1.8.5 A Model of the Signaling Mechanism

Figure 1.11 shows the structure of the sensory domain of the *E. coli* aspartate receptor Tar. The domain forms a dimer that contains a composite coiled coil: the two central helices form a dimer-stabilizing two-helix coil, but each also participates in a three helix coil with two other helices in the same monomer [93]. This coiled coil structure leaves a fourth "signaling" helix in each monomer free to move in response to aspartate binding. The signaling helix is at the C-terminus of the sensory domain, directly upstream of the transmembrane helix (TM2) that connects to the HAMP and signaling domains in the cytoplasm. The Tar sensory domain has been crystallized in both ligand-bound and ligand-free states. Alignment of the two structures combined with other data

indicates that the signaling helix moves downward like a piston by 1.6±0.2Å upon ligand-binding [50,113].

The piston model is strongly supported by recent work in which membrane-proximal aromatic residues in TM2 were relocated up to three residues up and down the helix [114]. These residues anchor the helix at the membrane interface, so repositioning them mimics the piston motion inferred to occur upon ligand binding. Relocating the membrane-anchoring residues by one or two positions was compensated by methylation in the cytoplasmic domain that reset the receptor to its baseline kinase activation state and allowed cells to perform chemotaxis successfully. Larger dislocations of the membrane-anchoring residues could not be compensated by methylation, and created smooth-swimming or tumbly phenotypes consistent with the piston model hypothesis that downward motion of TM2 activates the kinase, causing a decrease in [CheY~P] that leads to tumbling, and upward motion of TM2 de-activates the kinase, causing an increase in [CheY~P] that leads to smooth swimming [114].

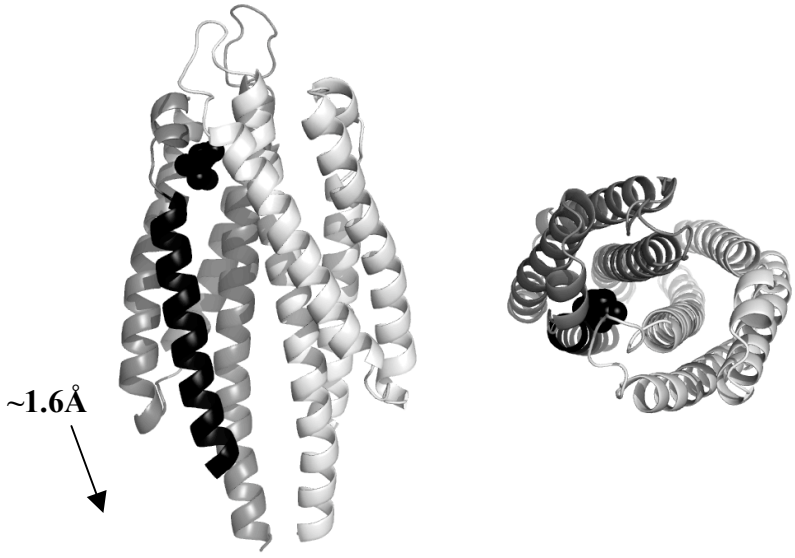

Figure 1.11 Piston signaling mechanism in the Tar sensory domain. (left) Side view and(right) top view of the periplasmic sensory domain of the *E. coli* Tar receptor (PDB code 2LIG). Bound aspartate is shown as a cluster of black spheres. The domain is a dimer; monomers are colored white and grey to differentiate them. The signaling helix in one monomer is colored black. Aspartate binding is thought to induce a downward motion of the signaling helix of 1.6±0.2Å [50,113].

For a long time, lack of structure in the HAMP domain was a bottleneck in understanding how the piston motion in helix TM2 might be transmitted to the MCP cytoplasmic domain and the kinase. The recent publication of a HAMP domain structure [107] and the proposal that signaling through the HAMP domain occurs by a concerted rotation of the helices in its parallel four helical bundle (Figure 1.9) raises the question of whether a piston motion from the sensory domain can induce rotation and change in coiled coil packing in the HAMP domain. A model signaling mechanism for receptors that share the membrane topology of Tar where a piston motion in the periplasm induces a rotation in the HAMP domain, which induces a change in supercoiling of the MCP cytoplasmic domain, is an attractive reference hypothesis (see section 3.3.1).

The authors of the HAMP domain model think that a piston motion is not compatible with their HAMP rotation mechanism, because their NMR data show that the register between helices in the bundle does not change in an activated mutant compared to the wild-type [107]. A piston motion would generate such a change in helix register. On the other side, cysteine scanning mutagenesis experiments in Tar show that signaling can occur in the presence of helix-joining disulfide bridges just below the HAMP domain [115], which appears to preclude a large rotational motion in the HAMP domain during signaling. The structure of the HAMP domain was determined from a single domain protein of unknown function [107], from *Archaeoglobus fulgidus*, a hyperthermophilic species with particularly thermostable proteins [116]. These unique characteristics may mean that its signaling mechanism is not generally applicable to other HAMP domains.

Indeed, experimental work in a variety of systems suggests that the dynamics and signaling mechanism in the HAMP domain may not be conserved across functional categories. In the sensory class II aerotaxis receptor of *E. coli*, Aer, the HAMP AS2 helix cooperates with the cytoplasmic PAS sensory domain to bind an FAD co-factor [102]. The HtrII receptor from the archaeal species *Natronomonas pharaonis* responds to light by interacting with a cognate, membrane-bound sensory rhodopsin. Signaling in HtrII has

been proposed to involve a rotation of the TM2 helix by 15° [117], which seems consistent with the HAMP rotation model. However, HtrII actually has two HAMP domains, and ironically, the first HAMP domain has been predicted to dissociate after TM2 rotation and interact with the rhodopsin [105]. These and other examples [85,96-106] show that HAMP function is not simple and appears to be malleable over evolutionary time when placed in different functional contexts.

1.8.6 Differences in Receptor Wiring

There are important differences in the way MCPs are wired for signaling in the two best-studied organisms, *E. coli* and *B. subtilis*. Recall that in *E. coli*, positive stimuli – addition of an attractant or removal of a repellent – inhibit kinase activity. Also in *E. coli*, methylation of any of four glutamate residues in the adaptation subdomain increases kinase activity. *B. subtilis* is wired differently with respect to both excitation and adaptation [118]. Positive stimuli in *B. subtilis* increase kinase activity. The overall behavior is the same – longer run duration in response to attractant – because both the receptors and the flagellar motor are wired oppositely. In *B. subtilis*, when the flagellar motor senses high levels of CheY~P, it continues turning counter-clockwise, extending the runtime [119,120]. In the McpB receptor of *B. subtilis*, different methylation sites have different effects on kinase activity. Methylating one residue increases kinase activity, but methylating another residue decreases kinase activity [121], in contrast to *E. coli* where methylation at any site increases kinase activity. These differences in both excitation and adaptation in the two best-studied chemotactic organisms are important to keep in mind as we analyze receptor diversity and its impact on evolution of the chemotaxis network architecture.

CHAPTER 2

MATERIALS AND METHODS

2.1 Databases

2.1.1 Sequence Databases

The Reference Sequences (Refseq) database at the National Center for Biotechnology Information (NCBI) is the ultimate source of all protein sequence data analyzed in this research [122,123]. Refseq is a curated subset of the Genbank database at NCBI. Each species catalogued by NCBI has a unique taxonomy ID number; organisms with sequenced genomes have separate Refseq accession numbers for each component of the genome. For example, *Vibrio cholerae* (NCBI taxonomy ID 243277) has two circular chromosomes with Refseq accession numbers NC_002505 and NC_002506. Essential information in the Refseq file for each component includes its full DNA sequence and the location, DNA sequence, and amino acid translation of each gene identified in the component. The core dataset used in this research consists of 236 complete and 76 draft prokaryotic genome sequences from Refseq (Table A.1).

The data in Genbank is mirrored by the DNA Database of Japan (DDBJ) and the European Molecular Biology Laboratory (EMBL). An important counterpart to Refseq is Swissprot, the protein sequence database curated by EMBL [124], and its uncurated companion database TrEMBL. Many global databases (for example, SMART) refer to Swissprot protein accessions, which are distinct from their Refseq accessions.

The core dataset was not accessed directly from NCBI, but instead from a developmental version of the Microbial Signal Transduction (MiST) database developed by Luke Ulrich [125,126]. The primary benefit of the MiST database is that it determines

the domain structure of all protein sequences in all genomes by scanning each one against the Pfam [127] and SMART [128] domain databases (see section 2.1.2), which is a computationally intensive task that requires access to a computer cluster. In the public or production version, the data in MiST is accessible only through a web interface, but the developmental version provides robust and versatile access to its Relational Database Management System (RDBMS) foundation via the Structured Query Language (SQL). MiST was first built on the MySQL RDBMS platform (http://www.mysql.com) because of its speed; later it was changed to PostgreSQL (http://www.postgresql.org) because of its more powerful database management and structuring tools [126].

Domain information can also be accessed directly from NCBI by intelligent use of the Entrez and Conserved Domain databases and the NCBI e-Utils programming interface [122]. There are currently 566 complete prokaryotic genomes available at NCBI [129]. In Chapters 5 and 6, I will outline preliminary work to automate the analysis outlined in this thesis so that chemotaxis pathways in newly sequenced genomes can be characterized automatically and integrated into MiST.

2.1.2 Domain Databases

Protein sequence data is much more useful and rich with information after it has been partitioned into families related by structure, function, and evolution. Generating a multiple sequence alignment of a protein family can be time-consuming and often involves significant expert knowledge of the family. A major product of this research, for example, is an improved alignment of the MCP cytoplasmic domain. Databases of domain models allow curated alignments to be distributed to and used by the scientific community. The two major domain model databases are the Pfam (Protein Families) database, first developed in 1997 [130], and the SMART (Simple Modular Architecture Research Tool) database, also first available ten years ago [131]. Pfam is updated much more frequently than SMART; the 312 genomes in this study were scanned against Pfam

17. The latest version is Pfam 22 [127], but chemotaxis protein models have not changed significantly since version 17. SMART is a more focused collection than Pfam; it specializes in models of signaling domains, since they are often more difficult to generate. The current version of SMART is 5.0 [128], and the genomes in this study were scanned against the latest model.

An important secondary source of domain information is the Conserved Domain Database (CDD) [122] at NCBI. CDD applies models from Pfam, SMART, and the Clusters of Orthologous Groups (COG) database [132,133] to all the sequence information stored at NCBI. CDD is important because its results are linked to all the other databases in the NCBI Entrez schema in a way that can be usefully queried at the NCBI website and using the NCBI e-Utils programming interface.

2.1.3 Structure Database – the Protein Data Bank

Models of the three-dimensional structure of proteins based on x-ray crystallography and nuclear magnetic resonance (NMR) experiments are stored in the Protein Data Bank (PDB). The PDB was established in 1971 at Brookhaven National Laboratory, but has recently been reorganized into a distributed world-wide database modeled after the successful collaboration between global sequence databases [134]. In this research project, analysis of PDB structures was performed with Pymol software (http://www.pymol.org).

2.2 Pairwise Sequence Alignment

A central issue in comparative genomics is the relationship between homology and sequence similarity. Two proteins are homologs if they perform similar functions and have shared ancestry; the presumption is that they were originally the same protein in an ancestral organism and have diverged by gene duplication (paralogs) or speciation (orthologs). Given the huge amount of genome sequence information now available,

establishing homology between sets of proteins is of fundamental importance. Homology is difficult to prove unequivocally, but a measure commonly used is sequence similarity. Two proteins that have high sequence similarity are assumed to be homologs.

Measuring sequence similarity depends on an amino acid substitution matrix consisting of the probabilities that each pair of amino acids will be found in the same position of two homologous sequences. Over short evolutionary times, amino acids separated by single mutations in their genetic code might replace each other. Over longer evolutionary times, protein structural requirements are more important, so amino acids with similar physical properties have high substitution probabilities. The two most popular substitution matrices, both of which are based on curated sequence alignments from the early days of protein sequence analysis, are PAM [135] and BLOSUM [136]. The BLOSUM62 matrix, built from a block of proteins at least 62% identical in sequence, is the default matrix used in ClustalW and BLAST and is perhaps the matrix in most widespread use today.

Besides point mutations, the other major process in the evolution of protein sequences is insertion and deletion of short segments; to account for such indels, sequence similarity measurements require a gap penalty to estimate the probability of gap insertion between two related sequences. While using the right substitution matrix is an important issue that does not get enough attention, the gap penalty and its effect on sequence alignment is a central concern of this research project, as will be outlined below.

The Needleman and Wunsch (NW) [137] algorithm generates a global alignment of the full length of two protein sequences. The algorithm involves placing the two sequences into a 2D grid or matrix, applying the scores from the substitution matrix and gap penalty to all possible pairs of residues in the sequences, and then calculating the optimal path through the grid, including paths with indels. Given a particular substitution matrix and gap penalty, this dynamic programming technique is guaranteed to find the

optimal global alignment between the two sequences. A minor change in the algorithm introduced by Smith and Waterman (SW) [138] finds the optimal local alignment. In a local alignment, the two most closely related pieces of sequence are aligned, neglecting the ends of both sequences if they contribute negatively to the score. Global alignment is generally restricted to comparing closely related sequences, whereas local alignment can, for example, compare two multi-domain proteins and find the one homologous domain they have in common.

2.3 Multiple Sequence Alignment

As the number of sequences to be compared increases beyond two, the computational complexity of the dynamic programming approach to sequence alignment becomes intractable. Progressive multiple sequence alignment (MSA) is a heuristic method that performs all possible pairwise alignments, then sorts and adds the pairs to the alignment in order of decreasing similarity, on the assumption that alignments built from closely related sequences are the most reliable [139]. The third sequence is aligned to the average profile of the first two, and so on until all sequences are aligned.

A problematic issue for progressive multiple sequence alignment is the correct placement of gaps. The gap penalties used in the pairwise alignment algorithms expect a geometric distribution of gap lengths; this fact is a by-product of the mathematical technique and bears little relation to gap lengths actually found in nature [140]. Alignment tools therefore use rules of thumb to guide the placement of gaps. ClustalW [141], the most popular first-generation MSA tool, uses heuristics meant to place gaps correctly in globular proteins. In an aqueous environment, a globular protein consists of a core of mostly hydrophobic residues surrounded by surface loops branching from the core that contain most of the hydrophilic residues in the protein. Over evolutionary time, gaps in globular proteins tend to occur in these surface loops. ClustalW takes advantage

of that fact, favoring the placement of gaps in hydrophilic regions. This heuristic is inappropriate for aligning coiled coil proteins like MCPs (see section 2.9).

2.4 BLAST

The BLAST algorithm (Basic Local Alignment Search Tool) [142] was developed to combat the problem of computational complexity faced by the sequence alignment methods based on dynamic programming. While dynamic programming can find the optimal alignment between two sequences, it does so essentially by generating all possible alignments and choosing the best. BLAST drastically reduces the search space by scanning for short "words" or k-mers of identical sequence, then restricting its search space to those areas where words cluster above some threshold level. It then uses dynamic programming to complete the alignment within the restricted search space. In the era of scanning millions of protein sequences to find homologs of a protein of interest, dynamic programming is intractable and BLAST is the only option. The simplest, most wide-spread way in comparative genomics to identify homologous proteins in two genomes is to look for reciprocal best BLAST hits.

2.5 Domain Architecture Prediction and Analysis

Multiple sequence alignments contain much more information than the set of protein sequences that compose them contain by themselves. Aligning a sequence against a pre-existing alignment is a much easier procedure than building an MSA from scratch, since the pre-existing alignment may contain significant manual editing based on human expertise that cannot yet be encapsulated algorithmically. Awareness of this fact has led to the development of databases of multiple sequence alignments (see section 2.1.2) so that the community can benefit from the hard work of others. Most multiple sequence alignments represent domains; a domain in a protein sequence can be defined in two ways. First, in terms of protein folding and structure, a domain is the set of protein

sequences that fold autonomously into the same three-dimensional structure. A good rule of thumb is that the average length of a domain is ~100 amino acids. Second, a domain can be defined as the set of proteins that make up a homologous family that performs a specific function. There is a significant amount of overlap between these two definitions, although technically a functional domain might be unstructured or consist of multiple folding domains.

2.5.1 PSI-BLAST

BLAST analysis is another example where using information from multiple sequences is more powerful than relying on a single sequence. A key step in multiple alignment is aligning a third sequence to the average of the first two; this is the simplest kind of profile alignment. The information in an MSA or profile can be formalized into a position specific scoring matrix (PSSM); a PSSM is essentially a statistical model of which amino acids are most likely to be found in specific columns of the alignment, as well as where gaps are most likely to occur. In Position-Specific-Iterative BLAST or PSI-BLAST [143-145], the results of an initial BLAST search that match the query sequence above some threshold value are aligned and a PSSM generated. Then another iteration of BLAST is run using the PSSM from the first iteration. This procedure can be iterated many times until convergence, when no more sequences are found that meet the threshold. At each iteration, the statistical power of the search is increased, as long as care is taken to exclude false positive matches that would skew the statistical profile incorrectly. PSI-BLAST analysis is a powerful tool for identifying new functional domains and protein families.

2.5.2 Hidden Markov Models

A Hidden Markov Model (HMM) [146-148], like a PSSM, is a statistical model of the information contained in a multiple sequence alignment. While PSSMs are

restricted to this context, HMMs can be applied much more widely; they are a general statistical tool for modeling many kinds of data. A significant portion of the statistics underlying HMMs, for example, was developed in the context of speech recognition [149]. HMMs are useful in many contexts within computational biology beyond modeling protein domains. For example, in the context of gene finding in eukaryotes, HMMs can be built to differentiate between exons, introns, and regulatory regions. The main weakness of HMMs is the assumption that the elements within the model are statistically uncorrelated; HMMs have no memory [148]. That assumption is strictly incorrect for protein domains, since adjacent positions in a protein structure interact with each other, but HMMs remain a powerful tool for protein domain analysis.

Within this thesis, HMMs are used extensively as tools for modeling and analyzing protein domains. The HMMer software package (http://hmmer.janelia.org/) is closely associated with the Pfam domain model database and was used for all of the HMM analysis in this thesis [127,146,147].

2.5.3 Computational Identification of Chemotaxis Proteins

Chemotaxis proteins were identified in the MiST database based on their characteristic combinations of Pfam and SMART domains (Figure 2.1). For each protein type, a SQL query to the database was constructed to find all proteins that included the domains in column 2 of Table 2.1 and excluded the domains in column 3 of Table 2.1. For example, while CheC has two copies of the CheC domain model from Pfam and no other domains, CheX has only one copy of the same domain model and no other domains. CheW has at least one copy of the CheW domain model from Pfam and no other domains. In order to differentiate it from CheA, CheV contains the CheW and Response_reg domain models from Pfam but excludes the HATPase_c domain model.

Not all of the models of chemotaxis domains in Pfam and SMART are of high quality. Through PSI-BLAST analysis, Kristin Wuichet has found CheC, CheD, CheX, and CheZ proteins that were not identified by the Pfam or SMART domain models [25]. Based on multiple sequence alignments from her PSI-BLAST results, I generated HMMs for each of these cases and scanned the genomes in MiST and Refseq for their presence. The results of this analysis are not yet included in the visualization of chemotaxis pathways in the Cheops database, which still relies on SQL queries to the pre-computed results from MiST (see chapters 5 and 6).

Table 2.1 Domain combinations used to identify chemotaxis proteins. SQL queries to the MiST database were generated that included the domains in column 2 and excluded the domains in column 3. All queries were based on the Pfam rather than the SMART domain model, since the models are of equivalent statistical power and coverage of all chemotaxis domains is better in Pfam.

Protein	Included Domains	Excluded Domains
CheA	CheW, HATPase_c	
CheB	CheB_methylest	
CheC	CheC, CheC	all other
CheX	CheC	all other
CheD	CheD	
CheR	CheR	
CheV	CheW, Response_reg	HATPase_c
CheW	CheW	all other
CheY	Response_reg	all other
CheZ	CheZ	
MCP	MCPsignal	

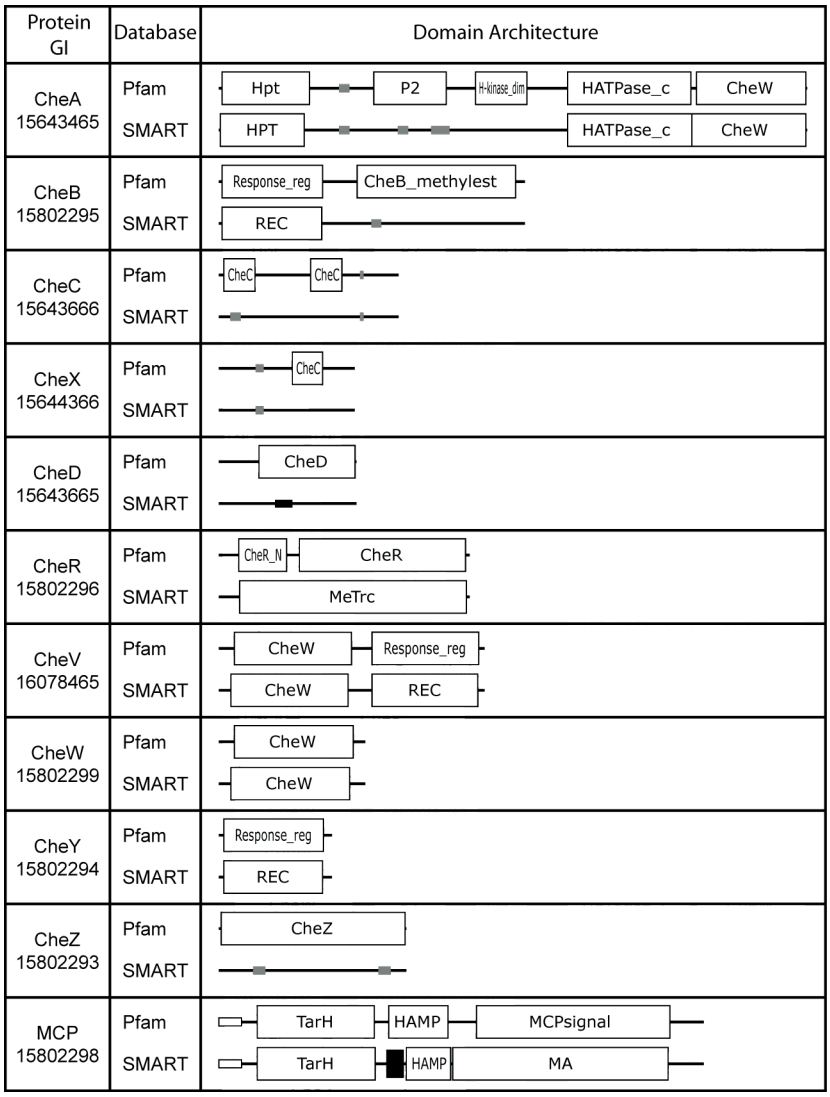

Figure 2.1 Domain architecture of chemotaxis proteins as visualized in MiST. The MiST database contains domain models from both the Pfam and SMART databases. Domains are shown as white boxes with their names inside. Small black, grey, and white boxes indicate predicted transmembrane, low complexity, and signal peptide regions, respectively. The NCBI database GI numbers corresponding to each protein sequence are given under their respective protein identifications.

2.5.4 Chemotaxis Gene Neighborhoods

It is a remarkable fact of biology that functionally related genes tend to cluster together in prokaryotic genomes [150], and that this clustering occurs to a lesser extent in eukaryotic genomes. Chemotaxis genes tend to cluster around the kinase CheA, so it is important to extract gene neighborhood information for chemotaxis proteins. Location along the chromosome is stored in the MiST database and in Refseq accession files, so it is an easy matter to generate gene neighborhood information. For each chemotaxis gene, the Cheops database includes information about other chemotaxis genes in the surrounding neighborhood of 30 genes. An improvement on gene neighborhood prediction would be to predict the actual operon structure around CheA [151]. For example, in *E. coli*, the chemotaxis genes in the neighborhood of CheA actually lie in two operons that are transcribed separately [67]. Operon structure has important implications for gene expression noise and how it is controlled (see section 1.6), and should be included in future versions of the Cheops database.

2.5.5 A Note about the HAMP Linker Domain Model

Later it will be important to locate HAMP domains in MCP sequences in order to define the N-terminal boundary of the MCP cytoplasmic domain. HAMP domain models, though, are imperfect. Both the Pfam and SMART HAMP domain models have low sensitivity. Each model picks up some HAMP domains that the other model misses (Figure 2.2). Finally, the model from SMART extends three residues farther at its C terminus than does the Pfam model (Figure 2.3). For the purposes of determining the N-terminal boundary of the MCP cytoplasmic domain, we used the Pfam model where possible and otherwise shortened the end of the SMART model by three residues.

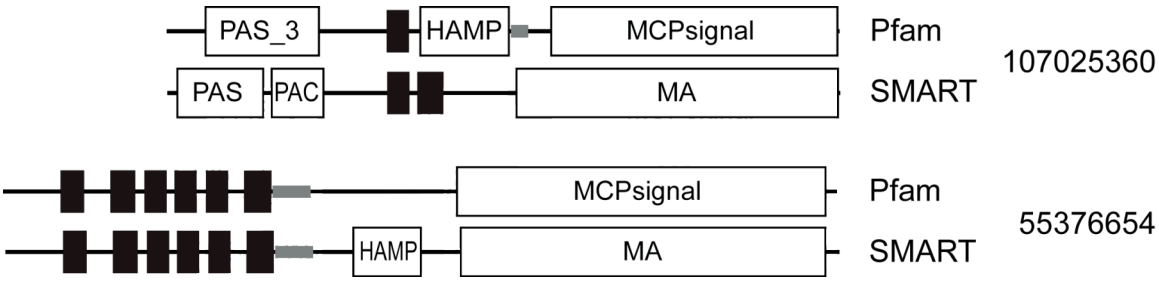

Figure 2.2 These two examples illustrate that neither the Pfam nor the SMART HAMP domain model is of high sensitivity. Each one identifies HAMP domains overlooked by the other. Figure elements as in Figure 2.1.

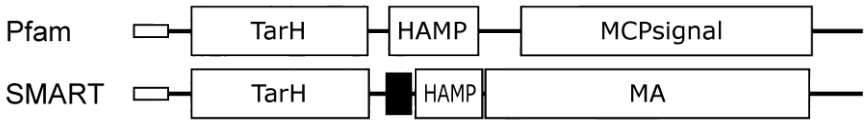

Figure 2.3 Differences in the Pfam and SMART HAMP domain models. The Pfam domain extends too far at its N-terminus, often overlapping a transmembrane region. The SMART domain model extends 3 amino acids further at its C-terminus than the Pfam model. Pictured is the Tar receptor from *E. coli*, Genbank ID 15802298. Figure elements as in Figure 2.1.

2.6 Analysis and Visualization of Sequence Conservation

The information content (IC) of each column in a multiple sequence alignment is defined by the formula

$$IC = \log_2 N - \left(- \sum_{i=1}^{N} p(x_i) \log_2 p(x_i) \right)$$

where N = 20 for protein alignments since there are 20 amino acids, and $p(x_i)$ is the frequency of amino acid x_i in the column [152-154]. For proteins, the maximum information content is $\log_2 20 = 4.32$ bits. It is sometimes easier, as in Figure 3.1, to think in terms of conservation level by normalizing IC (IC / IC_{max}) so that it ranges from 0 to 1, or least conserved to most conserved.

Sequence logos are an intuitive way to visualize the information content in multiple sequence alignments [153]. In a sequence logo, each column in the alignment is represented by the one-letter code of the residue types found most frequently in that column; the height of each letter is proportional to its information content. The sequence logos in this work were generated by the WebLogo Perl script [154]. The script was modified to group and color residue types into classes most often found in coiled coil proteins; the default color scheme is more applicable to globular proteins. Residue groups and their coloring are as follows: small (ASTG), green; hydrophobic (ILMV), black; aromatic (HFWY), yellow; negative (DE), red; polar (NQ), magenta; positive (KR), blue; special (CP), cyan.

2.7 Phylogenetic Analysis

All phylogenetic analysis in this research project was performed using version 3.1 of the Molecular Evolutionary Genetics Analysis (MEGA) software package [155]. Trees were exported from MEGA in the Newick standard format [156] for further analysis by custom Perl scripts. The primary method of tree construction was Neighbor joining (NJ) [157] because of its ability to generate reasonable trees for large numbers of input sequences rapidly. NJ is a distance method, using as input the matrix of pairwise distances between proteins generated during the process of multiple alignment. NJ starts with a star tree topology, then alters the topology two branches at a time, choosing at each step to coalesce the pair of branches that minimizes total branch length in the tree.

The number of possible tree topologies for a multiple alignment of N sequences is related to the factorial of N, so for more than a few sequences, it is impossible to search exhaustively for the optimal tree topology [156]. For large datasets, the NJ method represents the best compromise between speed and accuracy [158]. For smaller datasets, the maximum likelihood (ML) method is sometimes preferred [159], but ML is infeasible for datasets of more than several hundred sequences.

2.8 Perl Scripts

In this research project, I have relied heavily on the Perl programming language, writing over 350 scripts to process data generated by standard computational biology tools, perform analysis, generate figures, and interface with the MiST and Cheops databases. Unfortunately, including these scripts in an appendix of this thesis is not feasible; they would take up too much space after reformatting. In the hope that my effort at writing scripts might be useful for future Zhulin lab students and others, I have archived a subset of the scripts written over the lifetime of this research project at the Cheops database website, http://genomics.ornl.gov/cheops/.

2.9 MCP Alignment Method

2.9.1 Computational Identification of MCPs

Because the signaling subdomain within the MCP cytoplasmic domain is so well conserved, it is a simple matter to find all the MCPs in a genome sequence. The Pfam and SMART domain databases both have models of the MCP cytoplasmic domain (Pfam accession PF00015, MCPsignal; SMART accession SM00283, MA) which can find all MCPs in a genome using the HMMer software package [147]. A total of 2133 sequences matching the MCPsignal domain model from Pfam were found in 152 of the 312 organisms under study. Because of the variable length of the MCP cytoplasmic domain, the Pfam domain model does not depict the full-length domain correctly. Our main task, then, was to do a better job determining the boundaries of the domain and its subdomains. Because traditional multiple sequence alignment tools like ClustalW [141] do a poor job of aligning the gaps in coiled coil proteins (see section 2.3), we developed a method to cluster MCP sequences into length classes. Avoiding the difficult issue of gap placement made it possible to build alignments of individual length classes in ClustalW.

2.9.2 Determination of MCP Length

To determine correct domain boundaries for as many MCP sequences as possible, we searched for sequence features defining the start, center, and end of the cytoplasmic domain (Figure 2.4). To find the center of the domain, we generated an HMM of the signaling subdomain from a subset of 1232 MCP sequences and used it to locate the central glutamate residue in all 2133 MCP sequences. To find the start of the signaling domain, we queried the MiST database for the closest upstream element and defined the start of the signaling domain as the first amino acid residue before the center position but after the closest upstream element. This element was most often a HAMP domain (Pfam accession PF00672, SMART accession SM00304) or a transmembrane (TM) region, identified using Phobius [160]. The data from MiST on other upstream domains was

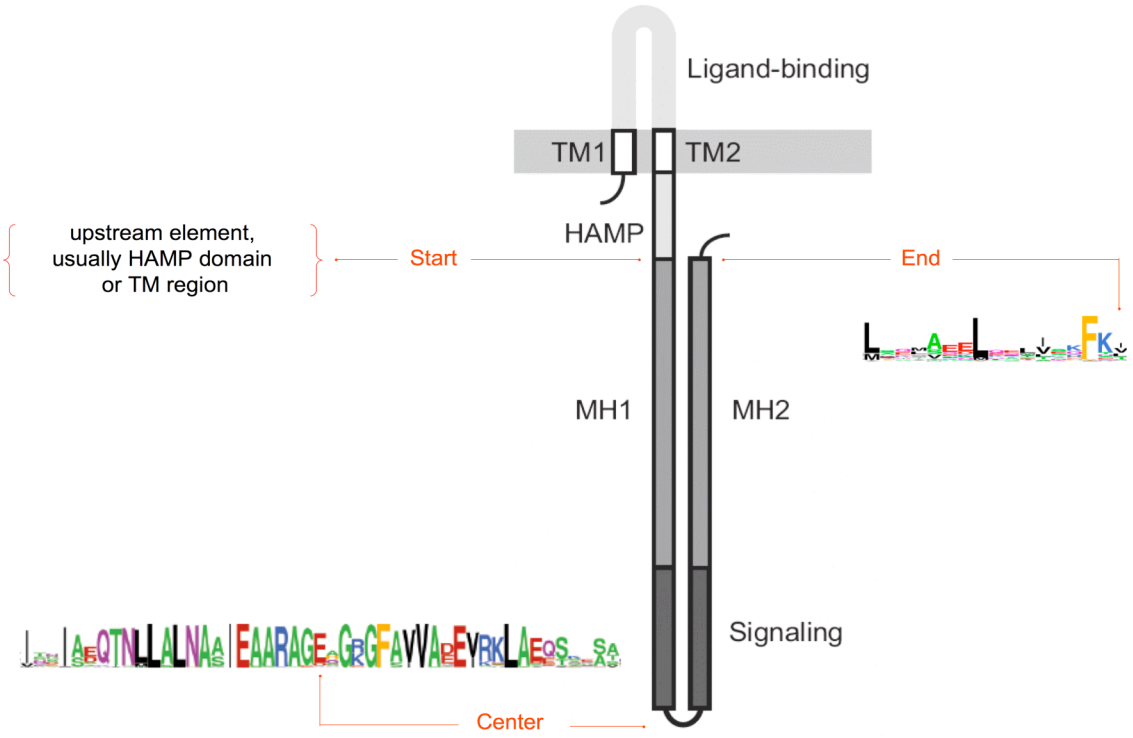

Figure 2.4 The lengths of the N- and C-terminal helical arms of the cytoplasmic domain were determined in each MCP sequence by finding the location of the start, center, and end of the domain using the indicated sequence features.

helpful, though, in some cytoplasmic MCPs that lacked both TM helices and HAMP domains. (See section 2.5.5 for a more detailed discussion of the HAMP domain model.) Where no upstream element was present, the first residue in the protein was identified as the start of the signaling domain. The end of the signaling domain was initially set to the C-terminal residue of the protein, but a preliminary alignment showed that many MCPs terminate with a conspicuous -L-x(6)-L-x(6)-F-x(2)- (LLF) motif (Figure A.1). Based on sequence logos, this motif was expanded to allow matches of [ILMVQ] at either of the two L residues. In sequences where such an LLF motif was found, the end of the signaling domain was redefined as the location of the nearest LLF motif after the center residue.

Defining start, center, and end positions allowed us to calculate the length of the N- and C-terminal helical arms of the domain. In previous work, LeMoual and Koshland found three classes of MCPs that differed in length by even multiples of seven residues [89]. We therefore selected the 1262 MCPs where the start-center and center-end distances differed by less than 7 residues and clustered them into 10 groups based on length. We named each length class based on the number of heptads it contained (Table 2.2); later, where possible, minor classes were renamed according to the major class from which they derived plus the number of inserted heptads (see section 0). We also grouped sequences based solely on the center-end distance, since the HAMP domain model is not strong and a missing HAMP domain generates a misleading start-center distance (see section 1.8.3). Center-end distance grouped another 199 MCPs into two additional length groups (38H and 28H).

Table 2.2 Distribution of MCP sequences across the 12 length classes at different stages of the alignment process

Alignment stage	Length class														
	All	Major								Minor*					
	Total	40H	36H	44H	38H	28H	34H	24H	Total	48H	38+4H 42H	40+12H 52H	38+20H 58H	40+24H 64H	Total
Initial	1461	531	344	237	148	51	43	36	1390	22	21	12	12	4	71
HMM Seed	1394	528	340	220	125	44	43	35	1335	19	18	9	9	4	59
HMM All	1846	671	490	294	158	48	59	47	1767	22	19	9	24	5	79
Manual	1915	694	496	331	158	48	61	47	1835	22	19	9	25	5	80
Gapless	1727	641	466	261	145	46	57	40	1656	20	19	9	18	5	71

*All classes were originally named based on the number of heptads (H) they contained; later, where possible, minor classes were renamed according to the major class from which they derived plus the number of inserted heptads.

2.9.3 Generation of Subfamily Hidden Markov Models

Alignments of each length class were generated in ClustalW 1.83 [141] with default settings and then manually edited by removing sequences that created gaps larger than one residue, trimming C-terminal tails past the LLF motif, and then trimming the N-terminus so that start-center and center-end distance were equal (Table 2.2). For each class a profile hidden Markov model was generated from the seed alignment using HMMer 2.3.2 [147] with default parameters. All MCP sequences were then scanned against the 12 domain models using the HMMpfam program from HMMer. A sequence was assigned to the top-scoring domain model if the model's score was at least 50 bits higher than the second-best-scoring model. This procedure resulted in the assignment of a length class to 1846 of 2125 MCPs; 8 sequences were discarded because all their bit scores were negative and they had poor matches to the HCD HMM. Of the remaining 279 MCPs, 69 were classified manually by examining the HMMpfam output, leaving 210 MCPs unaligned. All figures in this report were generated using only sequences that matched their top-scoring domain model without any gaps (Table 2.2), except that the methylation motifs in Figure 3.6 were generated from all categorized sequences.

2.9.4 Alignment of Subfamilies

The twelve individually aligned classes were merged into a single multiple alignment guided by the information content and amino acid consensus determined for each class, by profile-profile alignments, and by structural information from the Tsr [90] and TM1134 [94] crystal structures. Gap locations were positioned by aligning sequences from shorter length classes with those from longer classes following patterns of amino acid identity and similarity determined using sequence logos [154] and consensus scripts (available at http://coot.embl.de/Alignment/consensus.html). Strong conservation of knobs in the a and d heptad registers suggested that gap locations should be optimized by moving multiples of seven residues. Alignment editing was checked using pairwise profile-profile alignments in ClustalW 1.83 with default parameters. Most gaps from profile-profile alignments matched edited alignments within a few residues, so that repositioning of poorly conserved sites at the gap margins to maintain symmetry of gap location in the N- and C-terminal arms of the domain seemed acceptable. Classes 44H, 40H, and 36H correspond to Classes III, II, and I, respectively, from the earlier work of LeMoual and Koshland [89].

2.10 Analysis of MCP Structure

2.10.1 Coiled Coil Analysis

Knobs and holes were identified in the Tsr [90] and TM1143 [94] crystal structures using the SOCKET algorithm [92,93] with a 7.8 Å threshold. The SOCKET algorithm identifies coiled coils in protein structures by searching for knob residues all within a threshold distance of four hole residues on an adjacent helix, then looks for cycles of knobs to determine the number of helices in the coiled coil. Because alanine has

a short sidechain, alanine knobs tend to favor one triangular side or the other of the four-hole pocket. SOCKET fails to identify many alanine knobs because the distance between the knob and the fourth hole often exceeds the threshold.

Since knob residues were also identified from the MCP alignment by their heptad register, the set of hole residues associated with each knob was determined in the Tsr and TM1143 structures using the same distance measurement as in SOCKET and assuming the standard hole arrangement [92], except that three-residue holes were allowed for alanine knobs. The same hole residues were identified for all knobs found both by this technique and by SOCKET (data not shown).

2.10.2 Template Structures and Homology Modeling

Template structures of the 40H, 38H, 36H, 34H, 28H, and 24H classes were generated from the longer 44H TM1143 receptor crystal structure (PDB code 2CH7) by removing the appropriate residues and creating a new peptide bond between the appropriate backbone nitrogen (N) and carbon (C) atoms. If two sequences are related by the deletion of numbered residues n+1 to n+m as shown,

seq1: 1...n, n+1, ..., n+m, n+m+1,...L

seq2: 1...n, n+m+1,...L,

then the peptide bonds between N_n and C_{n+1} and between N_{n+m} and C_{n+m+1} in sequence 1 must be replaced by one peptide bond between N_n and C_{n+m+1} in sequence 2. These new peptide bonds were created in Pymol (http://www.pymol.org) with the *bond* command. To bring the two sequence fragments together and make the bond the correct length, the N, C_α, and C backbone atoms in a six-residue window around each new bond were aligned using the *pair_fit* command in Pymol. Table 2.3 shows which residues of TM1143 were deleted, according to the alignment, to create templates of each shorter MCP class.

Using the structural templates and the multiple sequence alignment of the MCP cytoplasmic domain, homology models of all categorized MCPs were built using Modeller 7.7 [161]. These models are useful for examining sequence features in a structural context (Figure 3.8), but require further refinement, like a round of energy minimization to remove TM1143-specific features of the coiled coil structure, before they could be used in molecular dynamics simulations.

Table 2.3 Residues deleted from the TM1143 crystal structure to generate templates of shorter classes

	Class					
Indel	40H	38H	36H	34H	28H	24H
1	248-261	309-329	248-261	248-261	248-261	222-291
2	489-502	422-442	329-342	323-343	304-345	461-529
3			409-428	406-447	406-447	
4			489-502	489-502	489-502	

2.11 Analysis of MCP Methylation Pattern

Methylation sites were identified in each length class by locating adjacent sites in the b and c heptad registers where the information content of glutamate or glutamine residues in the multiple sequence alignment exceeded 0.5 bits at both sites. To visualize methylation sites in different length classes, a Neighbor-Joining tree was constructed from ungapped aligned sequences using MEGA 3.1 [155] with the p-distance model of amino acid substitution and complete deletion of gap residues. The tree was partitioned into maximal subtrees containing receptors from just one length class. Methylation sites matching the consensus motif -[ASTG]-[ASTG]-x(2)-[EQ]-[EQ]-x(2)-[ASTG]-[ASTG]- were identified and visualized using custom Perl scripts. Sequences in Figure 3.6 were arranged in tree order, but with subtrees rearranged to cluster all receptors of the same class.

2.12 Determination of MCP Sensor Class

MCP sensor class and membrane topology can be easily determined by visual inspection of a two-dimensional domain model that includes transmembrane regions, as in Figure 1.7B. TM regions can be identified in MCPs and other proteins by a number of TM prediction programs. We tested Phobius [160] and DAS-TMfilter [162], then settled on Phobius because its predictions of TM regions in MCPs included fewer spurious predictions.

Using the data from Phobius on location and spacing of TM regions, the membrane topology and sensory class (Figure 1.7) of each MCP was determined as outlined in Table 2.4. Minimum domain size was set at 50 residues for this analysis. A total of 44 proteins had either 3 or 4 transmembrane regions, and were categorized manually. These data are available in the Cheops database.

Table 2.4 Distribution of Sensory Classes in MCPs

Sensor Class	Number	%	Rule for Determining Sensory Class from Transmembrane Topology and Spacing
I	1400	66%	2 TMs > 50aa apart
II	200	9%	1 or 2 TMs < 50aa apart, > 50aa from N-terminus
IIIc	164	8%	1 or 2 TMs < 50aa apart, <50 aa from N-terminus
IIIm	57	3%	> 4 TMs
IV	304	14%	no TMs
Total	2125		

CHAPTER 3

EVOLUTIONARY GENOMICS OF METHYL-ACCEPTING CHEMOTAXIS PROTEINS

3.1 Seven Major Length Classes

The alignment of the cytoplasmic domains of 1835 MCPs into seven major length classes (see section 2.9) results in a lot of data. A good overview of the data can be seen by plotting the level of conservation of each residue in the alignment (Figure 3.1A). The plot has two key features: first are the noticeable spikes of conservation throughout the alignment; these are the coiled-coil knobs in the a and d heptad registers. There are also noticeable subdomain features, namely the high conservation in the signaling subdomain and a medium level of conservation in the methylation helices generated by conserved methylation sites.

The two receptors in Figure 1.8A represent two of the seven major length classes. Tsr is 36 heptads long – 18 heptads for each of the two helical arms – and TM1143 is 44 heptads long, with helical arms 22 heptads in length. There are four gaps, two per helical arm, in Tsr relative to TM1143. Figure 3.1B shows a schematic alignment of the seven major length classes. Sequence logos of the alignment of the seven major length classes are in Figure A.1, and the full alignment is in Figure A.2. The length classes are named based on the number of heptads, so Tsr is in class 36H and TM1143 is in class 44H. An interesting feature of the alignment is that the gaps are all multiples of a heptad in length, and the gap locations are all symmetric about the center of the alignment. This pattern of gaps is very unusual and is the result of the unique structure of the MCP four helical bundle. If at some point over evolutionary time an MCP acquired an indel that was not a

heptad multiple in length, then the heptad register in the rest of the protein would become misaligned, the protein would not fold correctly, and that organism would not survive. Similarly, if a gap of the correct length arose in one of the helical arms without a gap of the same length appearing in the same location on the other arm, the register of the two arms would be misaligned, the protein would not fold correctly, and again the organism would not survive. In the evolutionary record we see only these unique symmetric gaps of heptad length that are allowed by the MCP's coiled coil structure (but see section 3.5). We feel sure that as more genomes are sequenced, more MCP length classes will be found, but we predict that all of them will have this symmetric gap structure. Such a unique gap structure happens very rarely over evolutionary time, so the MCP cytoplasmic domain is evolving very slowly.

3.2 Subdomain Boundaries and a New Subdomain

Using both the indel locations and the pattern of conservation throughout the domain, it is possible to determine subdomain boundaries precisely. Figure 3.2 is like Figure 3.1 except that subdomain boundaries based on these criteria have been indicated. A key at the bottom of the figure indicates that positions in the alignment are named based on their heptad register and number. The longest MCP class is 44 heptads long, so the numbering starts with heptad N22 at the top of the N-terminal helix and descends down to N01 at the base of the hairpin. The numbering continues with heptad C01 at the base of the C-terminal helix, and ascends up to heptad C22 at the top of the C-terminal helix. Residues within each heptad are labeled based on their heptad register. The advantage of this numbering system is that residues in the same number heptad on the two helical arms are physically near each other in the protein structure.

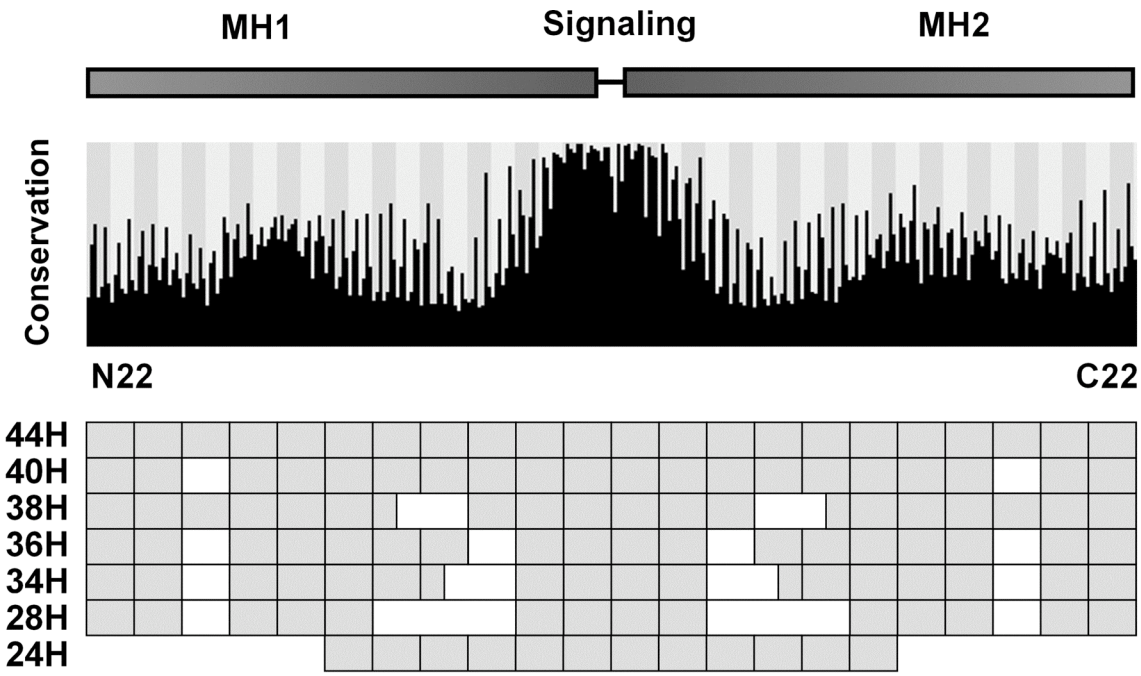

Figure 3.1 (A) Amino acid conservation within the MCP cytoplasmic domain. Position of each of 309 residues in the alignment is indicated by black columns. The column height shows the conservation level (see section 2.5.4). Position of each of the 44 seven-residue heptads (N22 to C22) is indicated by background grey shading. The first and fourth (a and d) knob residues of each heptad are strongly conserved. The schematic monomer above the conservation plot shows the poorly defined locations of the ethylation helices and signaling subdomain. (B) Schematic representation of the seven major length classes revealed by the multiple sequence alignment. Each rectangle represents a group of two heptads. Gap locations are shown in white. Subfamilies are named 44H through 24H indicating the length of the subfamily in heptads (H).

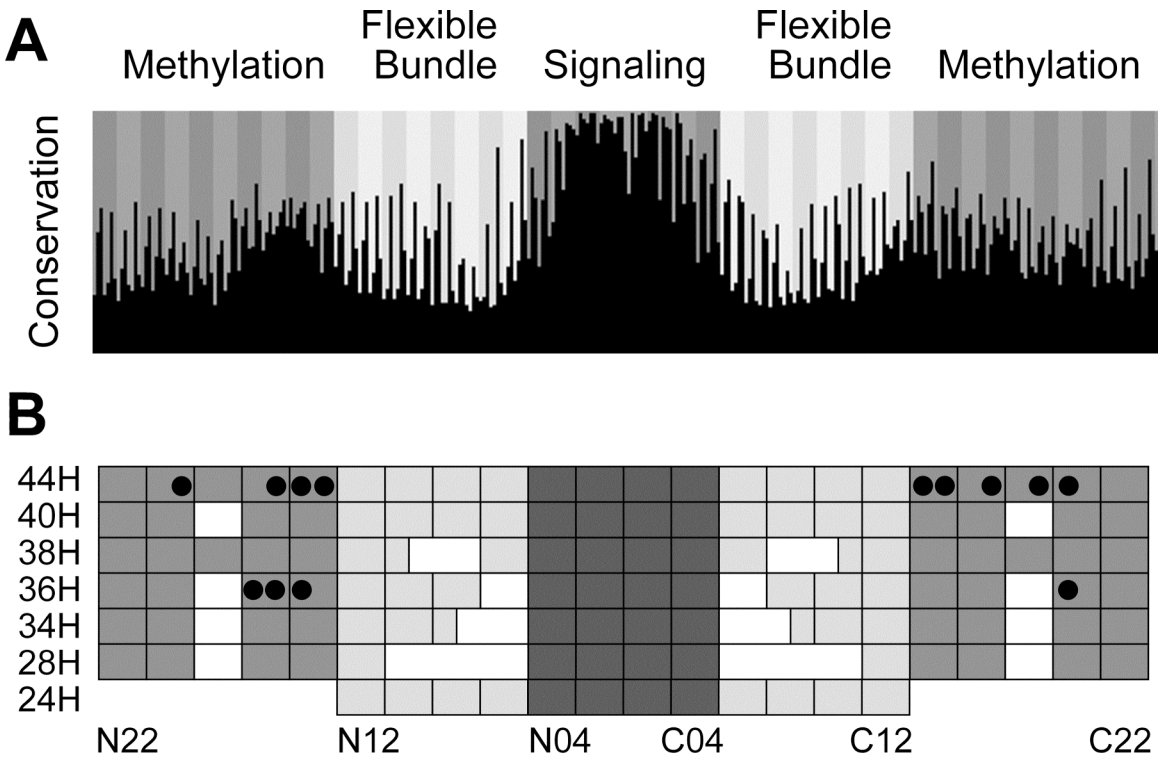

Figure 3.2 Subdomain structure of major domain classes. The three subdomains – methylation helices, flexible bundle, and signaling – are indicated by medium, light, and dark grey, respectively. (A) Amino acid conservation within the MCP cytoplasmic domain as in Figure 3.1. (B) Schematic representation of the seven major length classes. Each rectangle represents a group of two heptads. Gap locations are shown in white. Heptads are numbered from N22 at the N-terminus down to N01 at the center, and then up from C01 at the center to C22 at the C-terminus. This naming convention has the advantage that N- and C-terminal heptads with the same number are adjacent in the structure. Experimentally determined methylation sites in class 36H MCPs from *E. coli* [163] and class 44H MCPs from *B. subtilis* [76,121], *H. salinarum* [164,165] and *T. maritima* [81,166] are indicated by black circles.

We now define the exact boundaries of the signaling subdomain to be heptads 01-04 and of the methylation subdomain, heptads 13-22. The signaling and methylation subdomains are separated by two poorly conserved regions, consisting of heptads 05-12, where most of the gaps in the alignment are located. These regions have not been explicitly recognized in MCPs previously. We have termed these regions the flexible bundle subdomain.

3.3 Inferring Function from Sequence Features

Now we will examine specific conserved sequence features within each of the three subdomains and discuss their functional implications. The MCP cytoplasmic domain is extremely old; its divergence into subfamilies probably first occurred more than three billion years ago [167,168]. Sequence features in the domain that have been conserved by natural selection for that long must be functionally important. First we will examine the distinct pattern of knobs in the flexible bundle subdomain and discuss its implications for the signaling or excitation mechanism. Then we will examine the pattern of methylation sites in the methylation helices and discuss the adaptation mechanism. Finally we will examine the signaling subdomain and discuss receptor clustering and its implications for cooperativity and signal integration within the chemoreceptor array.

3.3.1 Signaling Mechanism in the Flexible Bundle Subdomain

The distinctive feature of the flexible bundle subdomain (FBS) is that only knobs are conserved there; it is therefore of interest to examine the FBS knobs more closely. Figure 3.3A shows a sequence logo of the FBS of one of the length classes, specifically 40H. In the logo, small residues (ASTG) are colored green and large hydrophobic residues (ILMV) are colored black (see section 2.6). Because only knob residues are conserved in the FBS, it is helpful to reduce the representation and look at a sequence logo consisting only of knobs (Figure 3.3C), with the knobs reoriented to match a

schematic diagram of the MCP monomer (Figure 3.3B). This reduced representation reveals an interesting pattern: in the N-terminal helix, starting at the top, there is a series of small knobs in heptads 10-12, followed by a series of mostly large knobs in heptads 05-09 near the base. The pattern is reversed in the C-terminal helix: in heptads 05-09 is a series of mostly small knobs, then a series of large knobs in heptads 10-12 at the top. The pattern at the base of both arms in heptads 05-09 is not as uniform as the pattern at the top.

We propose that the characteristic pattern of knob residue conservation in this subdomain is an important feature of the MCP signaling mechanism. We call the sections of helix with small knobs "tendons" and the sections of helix with large knobs "bones." The rigid bones are likely to stabilize the dimer structure while the flexible tendons are likely to transmit sensory information to the signaling subdomain. Recall that the a-d knob layers at the core of the MCP four helical bundle provides its stability (Figure 1.8). In the MCP dimer, matching residues in the two monomers are across each other on the knob layer diagonal. When there are large knobs on one diagonal of the knob layer and small knobs on the other diagonal, that square knob layer skews into a diamond shape (Figure 3.4A), because the large knobs attract each other more strongly via the hydrophobic effect than the small knobs. Many protein design experiments have been done on coiled-coils because they have a simple structure and folding mechanism. One group in particular built a four-helical bundle called the Alacoil with knob layers of alternating skew stacked one on top of the other [169]. The Alacoil is extremely stable, more stable than the average coiled coil. The pattern in the FBS is exactly the opposite – at the top of the bundle the bone and tendon helices create a series of knob layers all skewed in the same direction, with a transition between heptads 09 and 10, and at the bottom of the bundle, the knobs layers are skewed primarily in the other direction. Our hypothesis is that such a structure of stacked skewed knob layers is particularly unstable (Figure 3.4B).

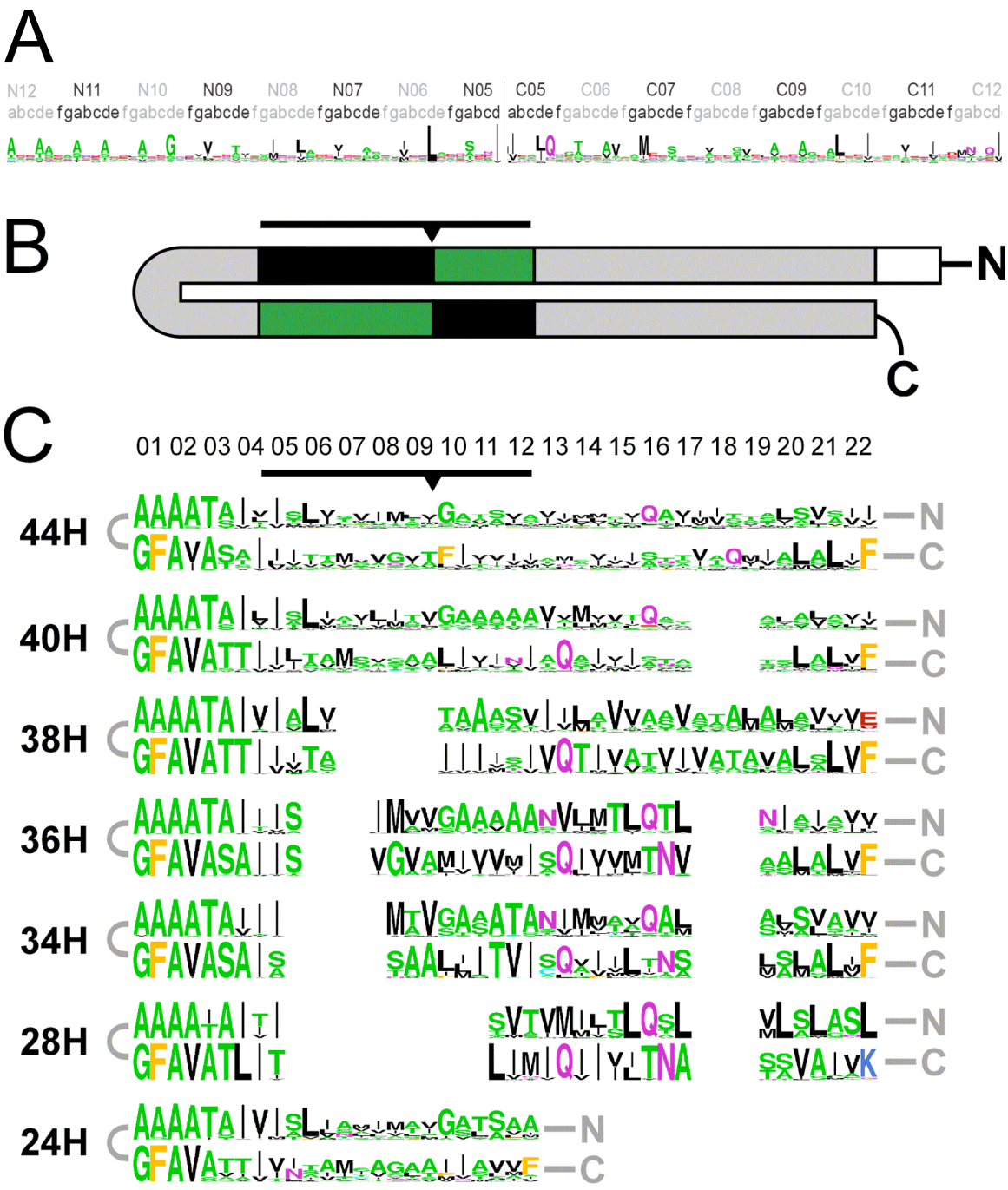

Figure 3.3 Knobs in the Flexible Bundle Subdomain. (A) Sequence logo of the class 40H FBS. (B) This schematic MCP monomer shows the orientation of the sequence logos in C. Bone and tendon helices in the FBS are colored black and green to emphasize that most knobs there are large and small, respectively. (C) Sequence logos containing just the knob residues from the seven major classes. Heptad numbering is indicated at top. Each column shows one a-d knob layer. There are two knobs per heptad. The FBS spans heptads 05-12 (black line). The glycine hinge between heptads 09 and 10 is indicated by a black triangle. Residues are colored as outlined in section 2.6.

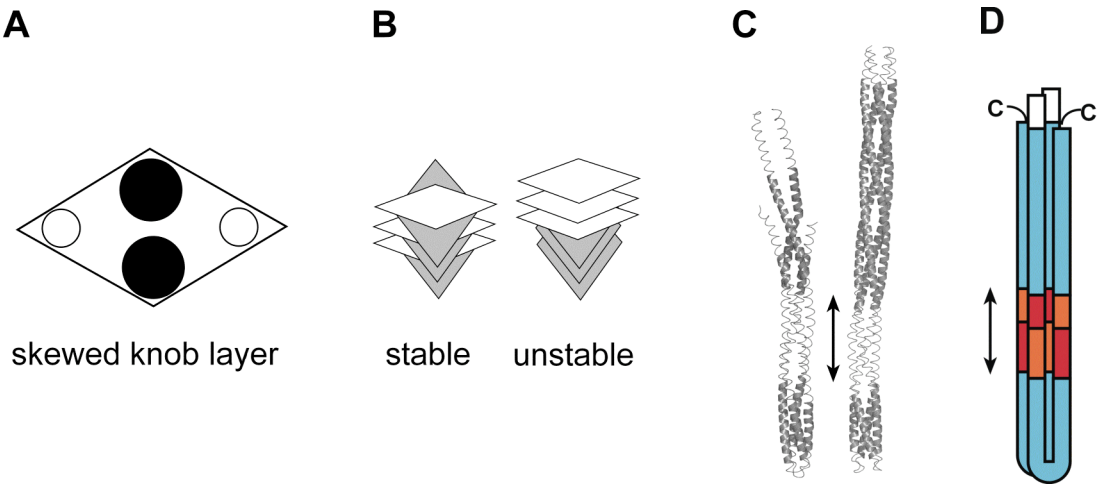

Figure 3.4 Function of knob layers in the FBS. (A) A skewed knob layer in a coiled coil. Layers formed from large (black) and small (white) knobs are skewed from square- to diamond-shaped. (B) Stable and unstable knob layers. Stacks of knob layers with opposing skew are stable, while stacks of knob layers skewed in the same direction are unstable. (C) Lack of a classical coiled coil in the flexible bundle subdomain. Structures of the Tsr (left) and TM1143 (right) signaling domain show coiled coil regions determined by the SOCKET algorithm with a 7.8 Å cutoff. Thin ribbons indicate regions where coiled coils were not detected. The bar indicates boundaries of the flexible bundle subdomain defined from the multiple sequence alignment. (D) Potential instability of the flexible bundle subdomain indicated by temperature factor. Colors indicate average temperature factor as follows. (i) Flexible bundle subdomain: "tendon" helices - red, very high; "bone" helices - orange, high; (ii) Methylation and signaling subdomains: blue, low. Detailed information is provided in Table 3.1. The bar indicates boundaries of the flexible bundle subdomain defined from the multiple sequence alignment.

Table 3.1 Temperature factors in the Tsr and TM1143 structures verify the functional importance of bone and tendon helices in the flexible bundle subdomain. Values are temperature factors averaged over all atoms in all residues from both monomers in the indicated region.

	Temperature Factor (Å2)		
Location	TM1143	Tsr QQQQ	Tsr QEQE*
FBS Tendons	76.5	94.1	148.9
FBS Bones	70.8	86.8	139.6
Signaling Subdomain	61.5	31.3	49.9
Methylation Subdomain	51.0	73.5	115.7

* Temperature factor data for the Tsr QEQE crystal structure was extracted from [170].

Scanning mutagenesis of glycine residues in the Tar receptor of *E. coli* revealed the importance of the transition point between heptads 09 and 10 [171]. Mutation of the highly conserved glycine knob at alignment position N10d to alanine or cysteine created a lock-on phenotype with constitutive kinase activation. Our analysis shows that this residue forms a conserved N10d / C10a knob layer, which is flanked above and below by bone and tendon helices in all major classes except 28H, where most of this subdomain has been deleted (Figure 3.3C). Coleman et al. [171] named this region the glycine hinge and argued that its function is to allow receptor dimers to bend, either to promote the initial assembly of the trimer of dimers (see section 3.3.3) or as part of the signaling mechanism. Conservation of a flexible bundle region with a central glycine hinge strongly supports the hypothesis that bending is central to the signaling mechanism.

Our results based on conserved sequence features are confirmed by analysis of the available structural data. Figure 3.4C shows the regions in the *E. coli* Tsr (class 36H) and *T. maritima* TM1143 (class 44H) structures where coiled coils were predicted using the SOCKET algorithm. SOCKET searches for knob residues within a threshold distance of four hole residues on an adjacent helix. SOCKET failed to identify the coiled coil structure of the FBS because the side-chains of small knobs there tend to favor one side or another of the four-hole pocket, so the fourth knob-hole distance often exceeds the threshold value. Early in the project, SOCKET analysis of the Tsr structure gave a clear visual indication that there was a region between the methylation and signaling subdomains with a unique structural and functional role.

Calculation of cross-diagonal distances in knob layers confirmed that the flexible bundle subdomains in both structures have two stacks of layers skewed in the same direction, whereas in the methylation and signaling subdomains there are stacks of alternating skewed layers (Table 3.2). The trend is not seen in the methylation subdomain of Tsr, because the crystal structure is disordered there and the four-helical bundle

Table 3.2 Diagonal distances across knob layers in the Tsr and TM1143 crystal structures. Positions of each residue in both the alignment and the crystal structures are indicated. Large knobs in each sequence are shaded black, small knobs grey. Distances between residues are calculated from the average location of all sidechain atoms in each residue, as in [92]. N and C subheadings indicate the diagonal distance between the same knob residue in the two N-terminal or C-terminal helices, respectively. The C - N column shows the difference between these two distances, which measures the degree of skew in the knob layer; negative values are shaded grey for contrast.

Alignment					Tsr							TM1143						
Heptad Number	Register N	Register C	Aln Position N	Aln Position C	Residue Type and Number N		Residue Type and Number C		Distances (Å) C-N	Distances (Å) N	Distances (Å) C	Residue Type and Number N		Residue Type and Number C		Distances (Å) C-N	Distances (Å) N	Distances (Å) C
22	a	d	3	307	VAL	267	PHE	515				LEU	224	TYR	528			
22	d	a	6	304	VAL	270	VAL	512				VAL	227	VAL	525	10.3	6.0	16.2
21	a	d	10	300	ALA	274	LEU	508				THR	231	LEU	521	-2.8	9.7	6.8
21	d	a	13	297	ILE	277	ALA	505				VAL	234	VAL	518	5.5	6.1	11.6
20	a	d	17	293	ALA	281	LEU	501				ILE	238	ILE	514	-2.1	8.8	6.7
20	d	a	20	290	ILE	284	ALA	498				ILE	241	VAL	511	6.6	5.4	12.0
19	a	d	24	286	ASN	288	SER	494				ASN	245	VAL	507	-1.8	8.2	6.4
19	d	a	27	283	-	-	-	-				ILE	248	THR	504	4.5	5.8	10.3
18	a	d	31	279	-	-	-	-				LEU	252	ILE	500	-2.8	9.1	6.4
18	d	a	34	276	-	-	-	-				ILE	255	LEU	497	1.2	6.9	8.1
17	a	d	38	272	-	-	-	-				MET	259	VAL	493	0.5	6.6	7.1
17	d	a	41	269	LEU	291	VAL	491				ILE	262	VAL	490	4.9	5.5	10.4
16	a	d	45	265	THR	295	ASN	487				ILE	266	ILE	486	-2.4	8.3	5.9
16	d	a	48	262	GLN	298	THR	484				ILE	269	ALA	483	4.6	5.6	10.2
15	a	d	52	258	LEU	302	MET	480	-11.5	17.0	5.5	VAL	273	ASN	479	-4.7	10.4	5.7
15	d	a	55	255	THR	305	VAL	477	-7.1	12.7	5.5	THR	276	MET	476	0.9	5.6	6.4
14	a	d	59	251	MET	309	VAL	473	-3.9	9.5	5.6	SER	280	ILE	472	-4.3	10.3	6.0
14	d	a	62	248	LEU	312	ILE	470	0.3	7.7	8.0	ILE	283	VAL	469	5.8	5.0	10.8
13	a	d	66	244	VAL	316	GLN	466	-4.1	9.8	5.7	THR	287	GLN	465	-4.8	9.7	4.9
13	d	a	69	241	ASN	319	SER	463	-0.1	7.9	7.8	ILE	290	ILE	462	6.6	5.0	11.5
12	a	d	73	237	ALA	323	ILE	459	-1.7	7.7	6.0	ALA	294	ILE	458	-2.6	8.6	6.0
12	d	a	76	234	ALA	326	MET	456	-2.7	7.9	5.2	ALA	297	LEU	455	-0.4	7.7	7.4
11	a	d	80	230	ALA	330	VAL	452	-4.2	8.8	4.6	ALA	301	ILE	451	-4.2	9.7	5.4
11	d	a	83	227	ALA	333	VAL	449	-0.9	7.0	6.1	SER	304	VAL	448	-0.1	7.2	7.2
10	a	d	87	223	ALA	337	ILE	445	-3.7	8.9	5.1	ALA	308	ILE	444	-4.8	10.3	5.4
10	d	a	90	220	GLY	340	MET	442	-4.0	9.8	5.8	ALA	311	LEU	441	-1.8	7.1	5.4
09	a	d	94	216	VAL	344	ALA	438	2.4	6.8	9.1	LEU	315	ALA	437	2.1	5.2	7.2
09	d	a	97	213	VAL	347	VAL	435	2.4	6.4	8.8	VAL	318	ALA	434	1.5	6.3	7.8
08	a	d	101	209	MET	351	GLY	431	6.7	5.5	12.2	THR	322	GLY	430	2.3	8.6	10.9
08	d	a	104	206	ILE	354	VAL	428	4.2	6.3	10.6	ILE	325	VAL	427	4.1	6.3	10.4
07	a	d	108	202	-	-	-	-	-	-	-	ALA	329	ILE	423	-1.5	8.5	7.1
07	d	a	111	199	-	-	-	-	-	-	-	VAL	332	SER	420	7.2	5.7	12.9
06	a	d	115	195	-	-	-	-	-	-	-	VAL	336	ALA	416	0.9	10.0	10.9
06	d	a	118	192	-	-	-	-	-	-	-	PHE	339	ALA	413	9.6	3.5	13.0
05	a	d	122	188	SER	358	SER	424	4.4	6.4	10.8	ALA	343	ILE	409	-6.4	13.7	7.3
05	d	a	125	185	ILE	361	ILE	421	2.0	6.9	8.9	ILE	346	VAL	406	4.4	5.8	10.2
04	a	d	129	181	ILE	365	ILE	417	0.0	7.2	7.2	VAL	350	VAL	402	-2.7	10.7	8.0
04	d	a	132	178	ILE	368	ALA	414	5.3	5.0	10.3	ILE	353	SER	399	5.0	5.3	10.2
03	a	d	136	174	ALA	372	SER	410	-4.4	10.2	5.8	ALA	357	SER	395	-3.1	9.7	6.6
03	d	a	139	171	THR	375	ALA	407	4.2	6.8	11.1	THR	360	ALA	392	3.7	5.9	9.6
02	a	d	143	167	ALA	379	VAL	403	-7.1	12.1	5.1	ALA	364	ILE	388	-4.3	9.1	4.8
02	d	a	146	164	ALA	382	ALA	400	3.7	7.7	11.4	ALA	367	ALA	385	1.0	7.7	8.7
01	a	d	150	160	ALA	386	PHE	396	-7.5	11.1	3.7	ALA	371	PHE	381	-3.2	8.5	5.3
01	d	a	153	157	ALA	389	GLY	393	3.2	5.9	9.1	ALA	374	GLY	378	1.8	7.8	9.5

structure is frayed at the ends. In knob layers of the TM1143 structure, one diagonal is on average 3.6 Å (34%) shorter than the other.

The temperature factor is a measurement of how much variability there is around the average location of each atom in a crystal structure. A higher temperature factor indicates greater flexibility. Averaging the temperature factor over all atoms in specific regions of interest reveals a pattern that confirms our sequence-based conclusions. In both the Tsr and TM1143 structures, the tendons have the highest temperature factor, and the bones have a lower temperature factor, but the temperature factor of the entire FBS is higher than that of both the methylation and signaling subdomains (Figure 3.4D and Table 3.1). The stacks of skewed knob layers do in fact generate greater flexibility in the tendon helices.

A popular model of the MCP signaling mechanism is the "frozen dynamics" model, which states that kinase activity is affected not by specific conformational changes in the MCP structure but by an overall change in structural rigidity throughout the domain [170]. Experimental support for the frozen dynamics model came from an unpublished crystal structure of the wild-type Tsr receptor. Tsr has four methylation sites, but in the wild-type, two of them are encoded as glutamines that are post-translationally deamidated by CheB (Figure 1.8A). The available structure of Tsr [90], PDB code 1QU7, is actually a mutant where the two methylation sites encoded as glutamate were mutated to glutamine to make the structure more stable and easier to crystallize. Wild-type Tsr is termed TsrQEQE and the mutant form is TsrQQQQ. Figure 2 of [170] shows that the temperature factor of TsrQQQQ is always lower than that of TsrQEQE; this evidence was seen as support for the frozen dynamics model, but the crystal structure of TsrQEQE was never deposited in the PDB.

Because we were interested in assessing the impact of methylation on the FBS, we digitally extracted the temperature factor data for TsrQEQE from Figure 2 of [170] and averaged it over the same structural regions as in the other two structures. Table 3.1

shows that the result was the same: the tendons had a higher temperature factor than the bones, and the FBS as a whole had a higher temperature factor than the rest of the domain. Even though the methylation sites are in the methylation helices, methylation changed the temperature factor more in the FBS than in the methylation subdomain, and more in the tendons than in the bones. The data thus support our conclusion that the MCP signaling mechanism travels through the FBS tendons.

3.3.2 Adaptation Mechanism in the Methylation Helices

The methylation subdomain (heptads 13-22) exists in all but one of the major MCP length classes. Remarkably, the entire methylation subdomain has been deleted from the 24H class (Figure 3.2 and Figure A.1). A striking feature of the alignment is the conservation of glutamate or glutamine (Glx) pairs that look like the consensus methylation sequence -[EQ]-[EQ]-x(2)-A-[ST]- found in *E. coli* MCPs [163] (Figure 3.5 and Figure A.1). Structurally, the Glx pair and small residues in the motif lie in adjacent turns on the solvent-exposed surface of a methylation helix. One residue of the Glx pair is the target for methylation by CheR and for deamidation and demethylation by CheB and CheD, while the flanking small residues are thought to be important for correct docking at the helix by the adaptation enzymes [46,172].

Figure 3.5 shows a sequence logo of several hundred receptors of class 36H. The methylation sites are known experimentally for only a few receptors in *E. coli*, but the sequence logo shows that their location and sequence are conserved in most receptors of that class. Methylation sites were identified in each length class by locating adjacent sites in the b and c heptad registers where the information content of glutamate or glutamine residues in the multiple sequence alignment exceeded 0.5 bits at both sites. To construct a consensus methylation motif for each class, we generated a sequence logo of a ten-residue window around the high-IC Glx pair sites (Figure 3.6, right). Merging class-specific information together resulted in a consensus methylation sequence for the

Figure 3.5 Conserved sites of methylation in class 36H MCPs. (A) Sequence logo from the 36H alignment of the four methylation sites known from experiments in *E. coli* [163], in heptads N16, N15, N14, and C19. Residues are colored as outlined in section 2.6. Red and green lines under the logo highlight the consensus motif (see text). (B) Location of the methylation site on the surface of the MCP four-helical bundle is indicated in this schematic diagram. The Glx pair (red) are in the b and c heptad registers and small residues (green) are in the e and f registers upstream and the f and g registers downstream of the Glx pair.

domain: -[ASTG]-[ASTG]-x(2)-[EQ]-[EQ]-x(2)-[ASTG]-[ASTG]-. This conservation pattern strongly supports the importance of small residues in the helical turns both upstream and downstream of the Glx pair. Sites matching this motif in all six classes are visualized in a dot plot of the alignment (Figure 3.6) and in structural models (Figure 3.7). It is evident that each signaling class has a different pattern of methylation sites, and these sites map onto different locations in the 3D structure of the signaling domain.

The conclusion from the correlation between length class and methylation pattern is that signaling and adaptation have co-evolved. Recall that *E. coli* and *B. subtilis* are wired differently for both signaling and adaptation (section 1.8.6). Adding attractant reduces kinase activity in *E. coli* but increases it in *B. subtilis*. Methylation of any site in *E. coli* increases kinase activity, but different sites have opposite effects in *B. subtilis*. Thus in the two best-studied cases, different signaling and adaptation mechanisms coincide with different length class and pattern of methylation. We have found five additional length classes, each with a unique pattern of methylation, that have not been experimentally studied. We expect this diversity in length and methylation pattern to be

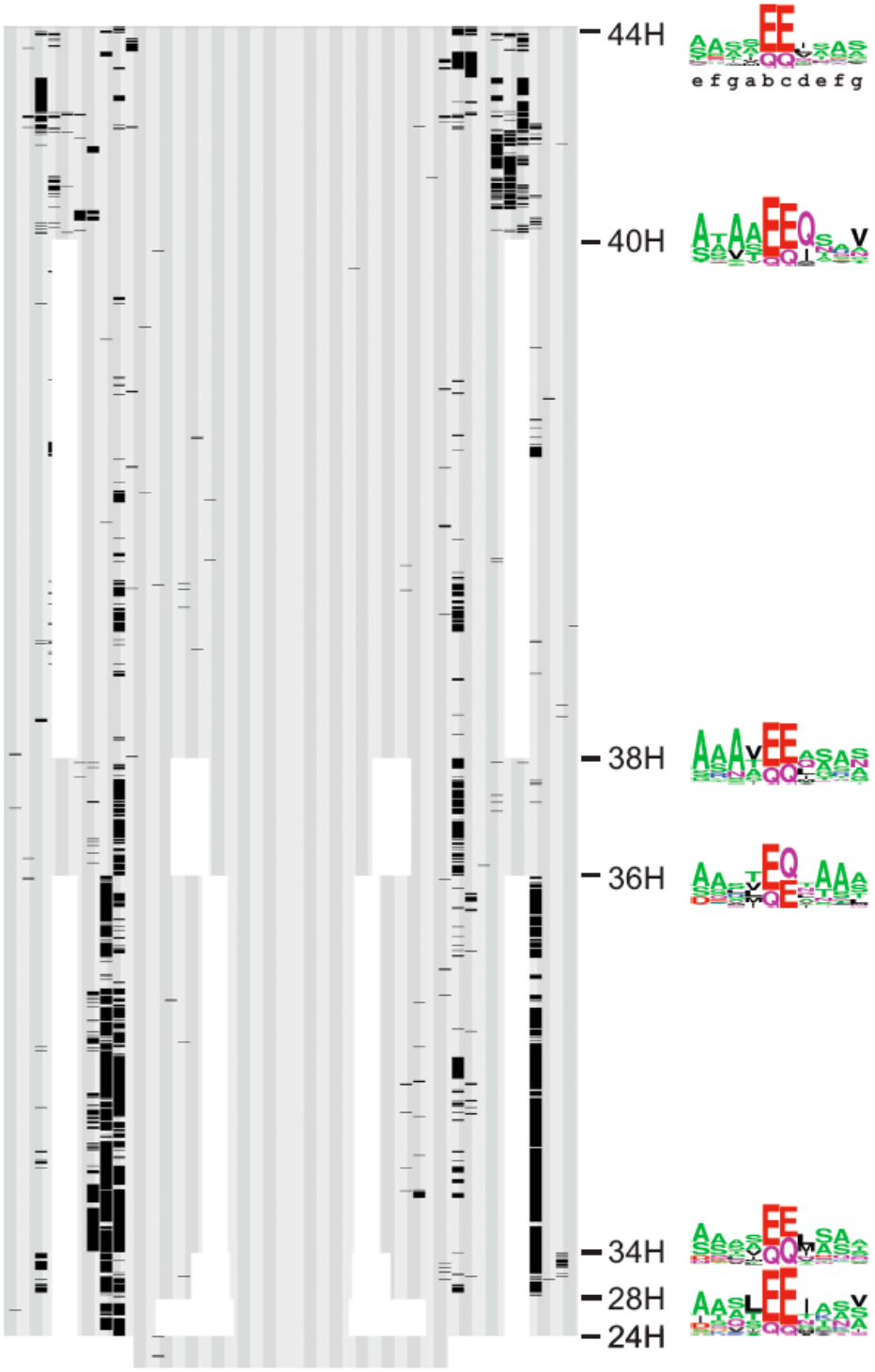

Figure 3.6 Methylation sites are conserved and located at class-specific positions. Dot plot of the MCP_CD alignment showing conserved predicted methylation sites (see section 2.11). A total of 1616 sequences are shown. Heptads (N22 to C22) are indicated by alternating grey shading. Positions of methylation sites matching the global consensus sequence are shown in black. Gaps are shown in white. Sequence logos of the methylation consensus sequence for each class are colored as outlined in section 2.6.

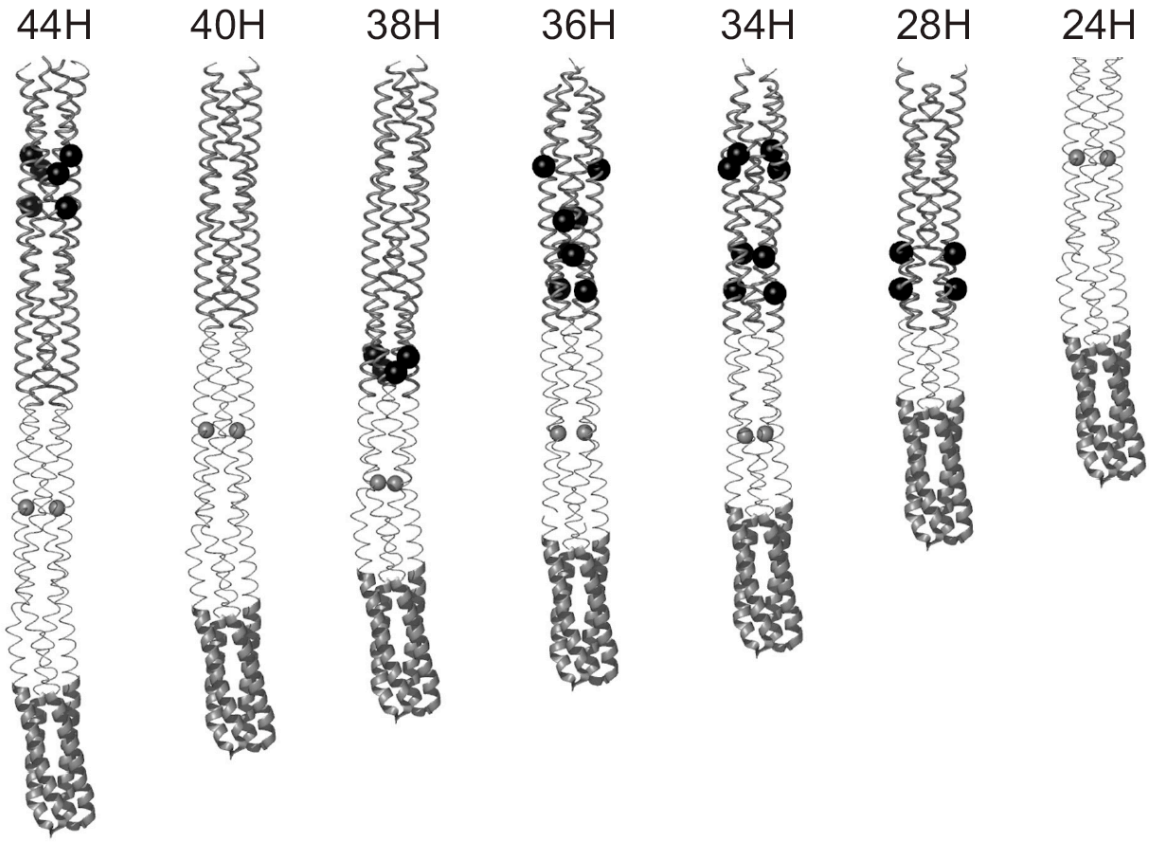

Figure 3.7 Template structures of major MCP_CD classes constructed from the *T. maritima* TM1143 structure (class 44H) show positions of the most common methylation sites (black spheres) in each class. The signaling subdomain is shown in dark thick ribbons, the flexible bundle subdomain in light thin ribbons and the methylation subdomain in dark thin ribbons. The glycine hinge is shown as a light grey sphere. Templates were constructed as described in section 2.10.2.

accompanied by differences in signaling and adaptation mechanism. It is interesting to note that one of the differentially methylated residues in *B. subtilis* is at alignment position C18c; this position is in a gap relative to the 36H receptors of *E. coli*. We see then a possible step change in evolution – assuming the 44H receptors are the ancestor of the 36H, loss of a methylation site understandably led to a change in adaptation mechanism.

The results outlined in Figure 3.6 and Figure 3.7 show large-scale methylation patterns across receptor classes. Looking more closely within each receptor class, there are in fact different methylation patterns among the receptors of individual organisms (Figure 3.8). These differences may play a role in mediating signal integration in the chemoreceptor array. Families of receptors with similar methylation sites may interact more closely within the array and sense suites of stimuli to which the organism needs to respond in a concerted way. Unfortunately, our knowledge of the sensory inputs of most receptors, from both experimental and computational analysis, is lacking (see section 1.8.4). In no organism other than *E. coli* has a complete census of sensory inputs yet been determined and mapped onto the MCP sensory domains. Perhaps the best progress towards this goal is being made in *Halobacter salinarum* [86,164,173-179]. Experimental inquiry into signal integration in chemotaxis is difficult. At the phenotypic level, competition experiments using swarm plate assays and release of caged compounds prove that signal integration does occur [28,30]. Molecular studies have focused on the interaction between the Tar and Tsr receptors in *E. coli* [34,35,180,181], but Figure 3.8A shows that these two receptors have the same sites of methylation, so these experiments do not probe the interaction between receptors with different methylation patterns. Progress in testing our hypothesis about the impact of methylation pattern on signal integration will rely on further developments in the field.

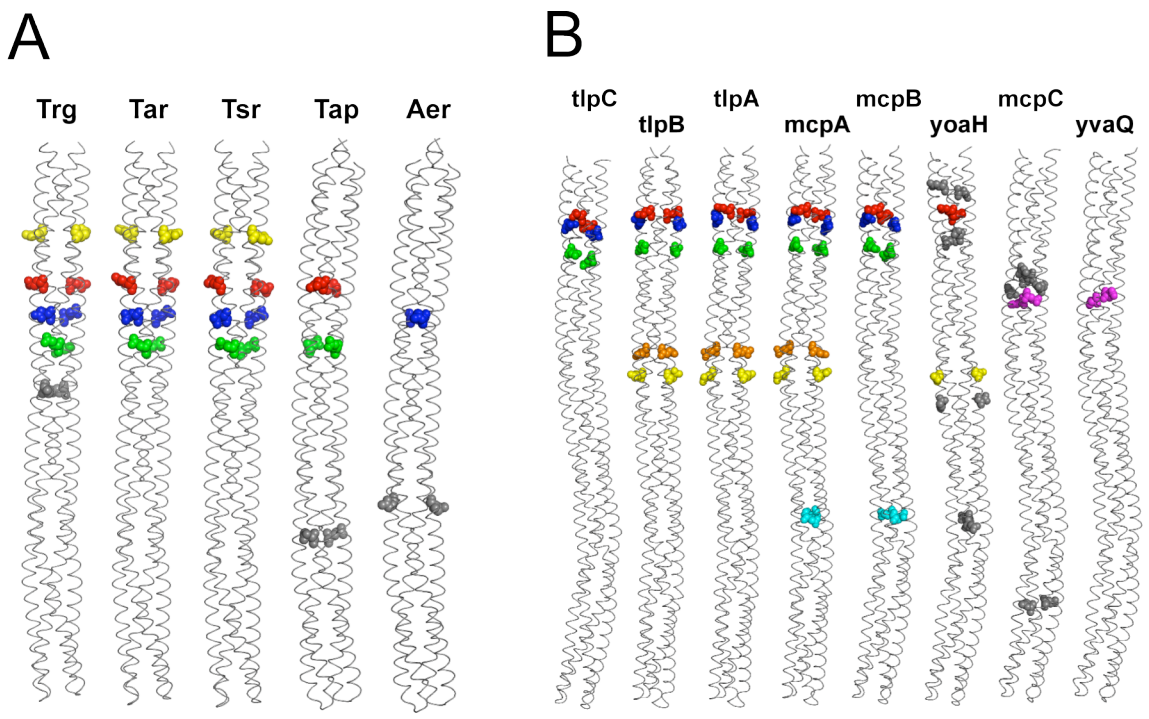

Figure 3.8 Diversity of methylation site pattern in (A) *E. coli* and (B) *B. subtilis* chemoreceptors. Mapping methylation sites onto homology models of the cytoplasmic domains of *E. coli* and *B. subtilis* MCPs reveals that there is diversity in methylation pattern even among MCPs of a single organism. These differences in methylation pattern may play a role in signal integration. The two cytoplasmic MCPs of *B. subtilis* are unaligned (see section 3.5) and are not pictured. Methylation sites are spheres colored according to how many receptors within the organism have a methylation motif at that site. From most to least frequent: red, green, blue, yellow, orange, cyan, magenta. Unique sites are grey. (A) and (B) are not to scale.

3.3.3 Receptor Clustering in the Signaling Subdomain

Most of the residues in the signaling subdomain are very conserved because of the multiple constraints of interaction with the scaffold and kinase proteins and with other MCP dimers. In this subdomain residues can be partitioned into two classes based on heptad register. Registers a, d, e, and g are intra-dimer contact sites. For example, position C02d of the alignment is a highly conserved valine residue in all seven major classes (Figure 3.9A). In the two known structures, that valine is located at the core of the dimer, mediating intra-dimer contacts and stabilizing the dimer (Figure 3.9B). Such strong conservation of the intra-dimer interaction sites suggests that all MCPs from

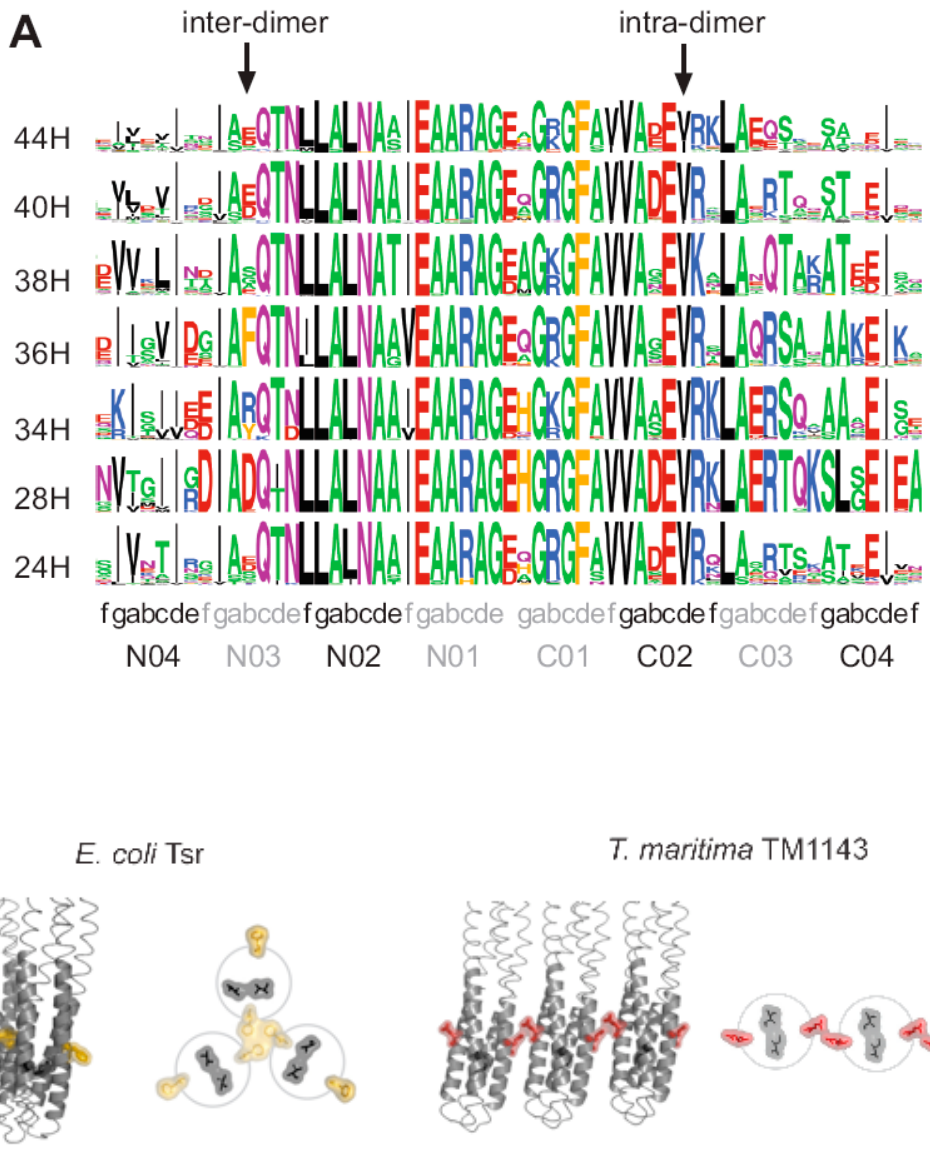

Figure 3.9 Family- and class-specific conservation in the signaling subdomain. (A) Sequence conservation in the seven major signaling classes. Two representative inter-dimer (N03b) and intra-dimer (C02d) interaction sites are indicated by arrows. (B) Visualization of selected inter-dimer and intra-dimer sites in the trimer of dimers of the *E. coli* Tsr and *T. maritima* TM1143 proteins. Inter-dimer, N03b: Phe (Tsr), Glu (TM1143); intra-dimer, C02d: Val (both structures). MCP dimers are outlined in grey to highlight their different organization in the two structures. Residues in the sequence logo and highlighted residues in the structures are colored as outlined in section 2.6.

different classes and organisms form dimers; however, there is no evidence for a preferential pattern of higher–order organization of dimers. The b, c, and f registers on the surface of the bundle are inter-dimer contact sites which mediate those higher-order interactions.

Interestingly, the most conspicuous class-specific residue in the signaling subdomain is a phenylalanine in the N03b position of the 36H class exemplified by the *E. coli* Tsr receptor (Figure 3.9B). This is a strong aromatic contact site of the trimer of dimers in *E. coli* [34,90]. On the other hand, this position in other MCP classes holds a strongly conserved charged residue and is consistent with the hedgerow organization of dimers of the 44H class receptor TM1143 [94]. Furthermore, a structural model of the CheA/CheW interaction in *T. maritima* suggests that a dimer, but not a trimer or dimers, can be accommodated by the scaffold-kinase complex in that organism [94]. Therefore, one possibility is that MCP dimers of different classes tend to form different patterns of higher-order organization. However, in cryo-electron microscopy studies in *E. coli*, three MCP dimers cluster together near CheA and CheW but do not form a trimer [182], suggesting that the model from *T. maritima* may apply more widely.

We propose that all these structural data are static snapshots of a conserved, dynamic signaling mechanism, a critical element of which is the oscillation of each MCP dimer between straight and bent conformations [183,184], one of which favors higher-order clustering more than the other. Recent analysis of the complex between CheW, CheA, and a straight receptor dimer in *T. maritima* suggests that the straight MCP dimer is in the kinase inactivating conformation [185]. If the signaling mechanism is conserved across species, this data from *T. maritima* is consistent with data from disulfide cross-linking experiments in *E. coli* that showed reduced cross-linking between MCP dimers after exposure of the array to the attractant aspartate [186]. According to the model,

straight dimers generate reduced kinase activity and tend not to associate with other dimers, while bent dimers activate the kinase and tend to form higher-order clusters. This "forest of dimers" model provides a structural basis for theoretical Ising models that explore how receptor cooperativity can enable high gain and wide dynamic range in chemotactic signaling [187,188]. MCP classes of different lengths and with different inter-dimer contact residues may have evolved as a way to tune the "infectivity" of ligand-binding, a key parameter of such models which quantifies how many neighboring dimers are affected by ligand-binding at one MCP dimer. Placing functional differences between receptor classes into an evolutionary context is an attractive feature of the model.

In the field of *E. coli* chemotaxis, the dominant paradigm is that trimers of dimers are stable and that signaling occurs within the context of the trimer of dimers [189] [190]. The "forest of dimers" model is based on the alternative paradigm that dimers are free to exchange between different trimers of dimers and mediate infectivity among more than two neighbors. It is possible that because of streamlining during the evolution of *E. coli*, trimers of dimers are in fact stable in *E. coli*, even though the ancestral chemotaxis mechanism was more dynamic. In this case, it is still possible to create a model for how receptor cooperativity has functioned throughout evolutionary history, except that for the special case of *E. coli*, inter-dimer interactions are so stabilizing that the trimer of dimers conformation has become frozen.

3.3.4 The Pentapeptide Tether

Perl regular expressions were used to scan all MCP sequences for the presence of the C-terminal pentapeptide tether to which CheB and CheR bind in *E. coli* [191]. The "assistance neighborhoods" of MCPs linked by this tethering mechanism have been found to enable precise adaptation in *E. coli* [192,193]. In *E. coli*, the pentapeptide motif is NWETF, but structural studies [45] and preliminary sequence analysis led us to expand

the motif to -x-[HFWY]-x(2)-[HFWY]-, allowing any aromatic residue in the second and fifth positions. The first step in the analysis was to determine the presence and length of any C-terminal extension by calculating the distance from the end of the MCP cytoplasmic domain to the C-terminus of the protein. One third of 2125 MCPs contain no C-terminal extension at all; another third contain an extension of less than ten residues. We found 217 MCPs in 67 of 152 genomes that do contain the consensus pentapeptide motif. These pentapeptide-containing MCPs are flagged in the Cheops database by a red marker (see chapter 5). All of these MCPs are of classes 36H and 34H, and all but two of the organisms where they are found are proteobacterial, implying that the pentapeptide tether is a recently-evolved mode of interaction between MCPs and adaptation enzymes.

3.4 Minor MCP Classes

Each of the five minor classes of MCP_CD is present in just one or a few related genomes (Table 3.3). All of them are derived from one of the major classes by a symmetric pair of heptad-length insertions in methylation helices 1 and 2. They are grouped in the schematic alignment of Figure 3.10 with the major class from which they derive and named according to difference in number of heptads from that class, except for class 48H. The originating major class most often yielded the second-best bit score in HMM searches of each minor class MCP sequence.

All of the aligned sequences in the spirochete *Treponema denticola* are of minor class 48H; other spirochete species, including *Treponema pallidum*, *Borrelia burgdorferi*, and *Borrelia garinii PBi*, have one or a few 48H MCPs while the rest of their MCPs are of class 34H. 48H MCPs are grouped in Figure 3.10 with class 36H, but may also have derived from class 44H, since the chemotaxis system of spirochetes is more closely related to that of *B. subtilis* than that of *E. coli*. 48H and 36H MCPs share a pair of gaps in heptads 05-07 relative to 44H MCPs, so two separate events may have generated a

Table 3.3 Minor classes of the MCP cytoplasmic domain. Number of sequences of each class and the name of the organism where the class is found are indicated. Appendix A contains two kinds of supplementary material: sequence alignments and sequence logo alignments of each parental class with its children. Taxonomy: dp, δ-proteobacteria; mp, magneto-proteobacteria; gr, Firmicutes; sp, Spirochetes.

Class	Number	Tax	Organism	Sequence Logo	Alignment
38+4H	19	dp	*Desulfotalea psychrophila*	Figure A.3	Figure A.4
38+20H	25	mp	*Magnetococcus sp. MC-1*	Figure A.3	Figure A.4
40+12H	9	gr	*Symbiobacterium thermophilum*	Figure A.5	Figure A.6
40+24H	5	dp	*Geobacter*	Figure A.5	Figure A.7
48H	20	sp	*Treponema denticola* et al.	Figure A.8	Figure A.9

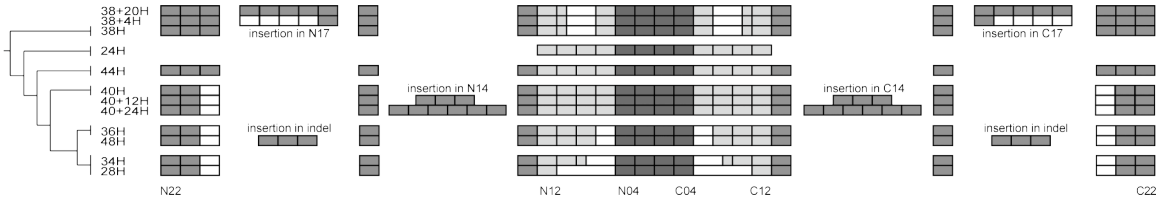

Figure 3.10 Subdomain structure of major and minor domain classes. All five of the minor classes derive from their parental class by symmetric, heptad-length insertions in the methylation subdomain. Minor class 48H is clustered here with 36H, but it may also be derived from class 44H. Figure elements are as described in Figure 3.2.

deletion in heptads 05-07 in classes 36H and 48H relative to the parent 44H class, making that location an evolutionary hotspot (Figure A.8).

Interestingly, the N-terminal insertion regions of the 38+20H and 38+4H classes appear to be related; they both contain the same strong signature of what may be a non-canonical methylation site (Figure 3.11). This fact appears to show that one of these two classes is derived from the other – 38+20H by insertion into an ancestral 38+4H sequence, or 38+4H by deletion from an ancestral 38+20H sequence – suggesting an evolutionary or ecological link between *D. psychrophila* and *Magnetococcus sp. MC-1* that would allow genetic exchange between them either by vertical or lateral transfer.

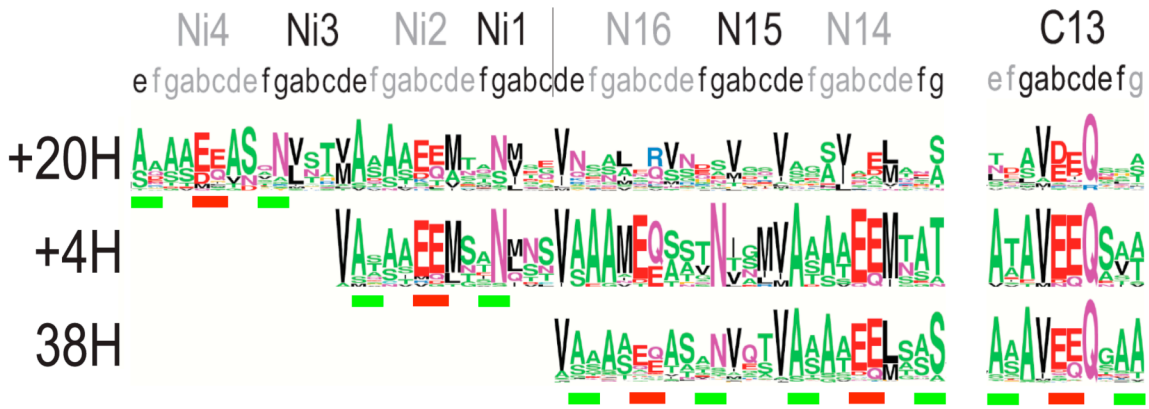

Figure 3.11 Methylation sites in class 38H and related minor class MCPs. Class 38H MCPs have methylation sites matching the global consensus motif in heptads N14 and C13. It also appears to have an unusual methylation site in heptad N16, where a C-terminal asparagine replaces the expected small residue. These three sites are all shared by the 38+4H MCPs in *D. psychrophila*, but not by the 38+12H MCPs in *Magnetococcus sp. MC-1*. The non-canonical methylation motif is present in heptad Ni2, an N-terminal insertion region shared by the two minor classes, and at an additional site, heptad Ni4, in the extended insertion region of the longer class. The inserted sequence shared by the two minor classes indicates that *D. psychrophila* and *Magnetococcus sp. MC-1* share ancestry or ecology and were not derived independently from the parental class.

Multiple sequence alignments of each minor class with its parent, as well as sequence logos of those alignments, are available in Appendix A; Table 3.3 lists the appropriate figure numbers for each case. It is curious that all of the minor classes derive from their parents by insertions in the methylation subdomain. The major impact of that fact is to verify that the MCP cytoplasmic domain can evolve not just by symmetric, heptad-length deletions from longer classes, but also by insertions into shorter ones. To date only 80 minor class sequences have been categorized. If more appear as new genomes are sequenced, it may be necessary to revise the heptad numbering scheme outlined in section 3.2. For now, heptads in insertion helices are numbered up from Ni1 and Ci1 starting at the base of the hairpin. All of the minor classes appear to be coiled coil insertions, except that for the longest insertion, in class 40+24H, there are still too few sequences to reveal the characteristic knob conservation pattern of a coiled coil.

3.5 Unaligned MCPs

After alignment of 1915 MCPs into seven major and five minor length classes, the remaining 210 sequences did not confidently match domain models of any class either due to truncation (30 sequences, mostly the result of genome assembly problems) or due to poor conservation and the presence of asymmetric indels. Aside from truncations, the remaining sequences were categorized, if not aligned, into three groups: Unaligned Cytoplasmic (UC), 121 sequences; Unaligned Membrane-bound (UM), 41 sequences, and Unaligned Other (UO), 18 sequences. Examples of the UC and UM classes are shown in Figure 3.12.

Both of the cytoplasmic MCPs from *B. subtilis* are UC sequences. The yfmS MCP is representative of most UC sequences; it is a stand-alone MCP_CD domain of unknown function. HemAT is an oxygen sensor, the sensory domain of which has been crystallized, although it is not clear how or whether HemAT interacts with the eight membrane-bound 44H MCPs in the *B. subtilis* chemoreceptor array [88]. The UM McpA receptor from *P. aeruginosa* resides in the gene neighborhood of the che2 kinase, but is in fact transcribed separately. Interestingly, fluorescent tagging has shown that McpA colocalizes not with che2, but with the che1 kinase in the major chemoreceptor array in *P. aeruginosa* [194]. All UM sequences share a characteristic membrane topology with one or two N-terminal TM helices and no identified sensory domains. We hypothesize that UM MCPs may play only a structural role in the array, anchoring other receptor, scaffold, and kinase proteins in the membrane.

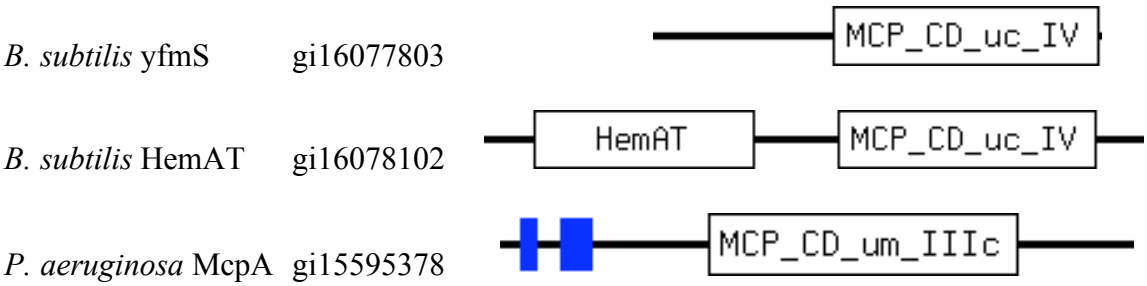

Figure 3.12 Examples of Unaligned Cytoplasmic (UC) and Unaligned Membrane-bound (UM) MCPs. Species, sequence name, and Genbank ID are indicated.

We have shown that symmetric pairs of indels are a key pathway of evolutionary change in the MCP cytoplasmic domain, but have not established a mechanism that can generate such unusual changes in sequence. We propose that cytoplasmic MCPs may provide a reservoir of decreased selective pressure where the maintenance of symmetry at the ends of the domain away from the hairpin turn is less important for structural stability and function than in their membrane-bound counterparts. Cytoplasmic MCPs in archaeal species, for example, fit the 44H class domain model less well than their membrane-bound counterparts because of gaps in this region (data not shown), and may represent an evolutionary bridge between class 44H and 40H receptors. This hypothesis could also explain the prevalence of UC MCPs with apparently asymmetric indel locations.

CHAPTER 4

COMPARATIVE GENOMICS OF CHEMOTAXIS

4.1 Kinase Diversity

Whereas receptor diversity is based on length class and methylation pattern, kinase diversity is based on gene neighborhood around and domain organization of the kinase CheA. Most of the chemotaxis proteins in any given genome, with the exception of the receptors, cluster in the gene neighborhood of CheA, probably forming operons. Interestingly, while some MCPs are in the neighborhood of CheA, most of them are not, and are instead scattered around the genome. CheA has five domains that are conserved to different degrees [27]. Hpt (P1) is the histidine phosphotransfer domain [195-197]; it contains the conserved histidine residue that gets phosphorylated by the kinase, and it interacts with and transfers the phosphate group to CheY. The P2 domain is a highly variable domain that interacts with receiver proteins to facilitate phosphotransfer [198,199]. P1 and P2 are connected by flexible linkers to a globular core of three conserved domains, the dimerization (P3), ATPase (P4), and CheW (P5) domains [200]. Some CheAs also have a CheY-like receiver domain (P6) fused to the kinase.

A multiple sequence alignment of the P3-P5 core forms a good basis for classifying CheA into subfamilies. Kristin Wuichet generated such an alignment and built a Maximum Likelihood phylogenetic tree from it [25] using the PhyML software package [159]. Figure 4.1 shows that tree with three kinds of data painted onto it: phylogenetic grouping, gene neighborhood, and domain organization. Although the tree is based on the P3-P5 alignment, there are clear correlations with gene neighborhood, phylogeny, and organization of all six domains, including P1, P2, and P6.

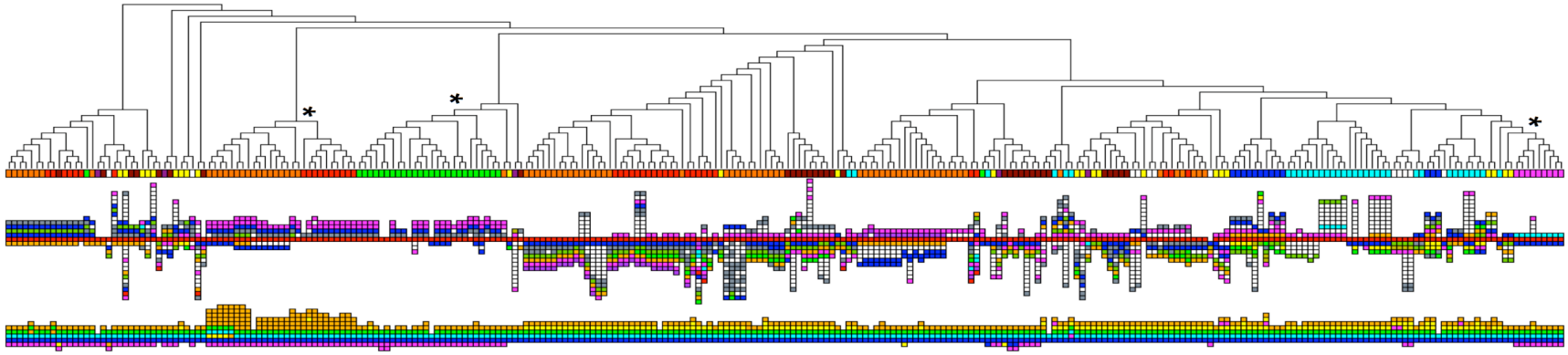

Figure 4.1 Maximum likelihood phylogenetic tree built from the three core conserved domains of CheA. Painted onto this tree are three additional types of information. From top to bottom, phylogeny, CheA gene neighborhood, and CheA domain structure. Phylogeny color scheme: archea, dark blue; gram-positive, cyan; cyanobacteria, green; proteobacteria: alpha-, maroon; beta-, red; gamma-, orange; delta-, yellow; epsilon-, magenta; magneto-, purple; other, white. Gene neighborhood color scheme (as in Figure 1.3 and Figure 1.6): MCP, grey; CheA, red; CheY, magenta; CheV, cyan; CheW, blue; CheB, orange; CheR, dark green; CheD, light green; CheC, yellow; CheX, pale yellow; CheZ, purple. CheA domain color scheme: P1 Hpt, orange; P2, yellow; P3 dimerization, green; P4 ATPase cyan; P5 CheW, dark blue; P6 receiver, magenta. Starred subtrees are shown in more detail in Figure 4.2.

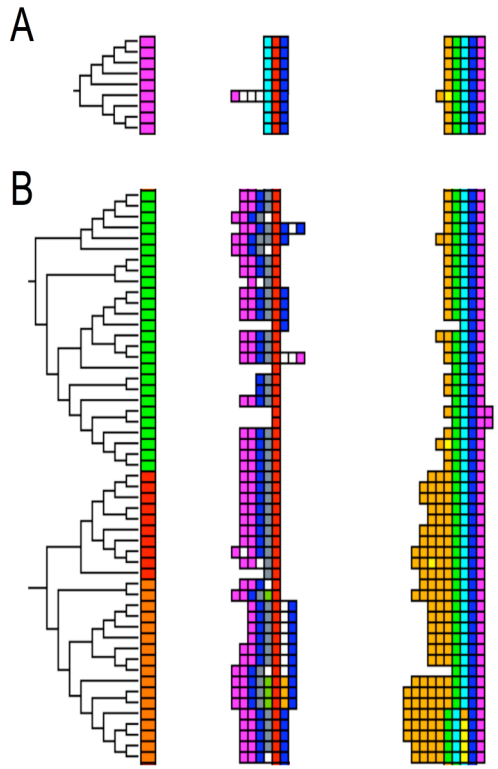

Figure 4.2 Detailed view of the CheA tree from Figure 4.1 showing (A) the F3 / VAW and (B) the Tfp / YWMA CheA types. Tree organization and coloring as in Figure 4.1.

Partitioning the CheA tree into subtrees and analyzing the phylogenetic diversity, domain organization, and gene neighborhood data in each subtree allowed a preliminary classification of CheA into subfamilies. Analyzing phylogenetic trees of accessory chemotaxis proteins and other information enabled Kristin to generate a more refined classification; more detail is available in her thesis [25]. Table 4.1 lists the preliminary designation of CheA types used in this project and shows their correlation to the types defined by Kristin.

Our original attempt to classify CheA based soley on gene neighborhood failed because operon structure evolves fairly rapidly in prokaryotes [150]. Gene neighborhood alone was useful in characterizing some classes, namely those with ARXY, VAW, YZAB, and YWMA neighborhoods. In those cases, the preliminary designation was

based on gene neighborhood. In other cases, gene neighborhood was much too diverse, but adding phylogeny created a sufficient basis for preliminary classification. These cases include the Ecoli designation, the CheA type found in *E. coli* and its relatives, the Grampos designation, a diverse class of CheAs found in gram-positives or Firmicutes, and the Alpha designation, a CheA type characteristic of α-proteobacteria. The Ecoli99 designation is uniquely based on the knowledge that this type of CheA interacts with class 34H MCPs (see below), which were called Class 99 MCPs in an earlier nomenclature. The preliminary designations YBACDR and WRWMAB are named based on CheA gene neighborhood, although some members of those classes have diverse CheA neighborhood and were included because of their phylogeny and position in the subtree. The "uncat" preliminary designations refer to small subtrees of the ML CheA tree that did not contain enough diagnostic information for useful characterization. Table 4.2 lists the same information as Table 4.1 re-ordered according to the final CheA designation. Both preliminary and final CheA designations are indicated in the Cheops database (see Chapter 5).

It is clear from Table 4.2 that the preliminary classification is less detailed than the final; the YBACDR and Grampos classes together make up the F1 system, which includes F1a and F1b subclasses that are undifferentiated in the preliminary scheme; likewise the Ecoli class matches the F7 system, but F7 is divided into F7a and F7b. The preliminary partition of CheA families does not include the F4 or F10 systems.

Table 4.1 CheA Classification. Phylogeny indicates most common phylogenetic groups containing CheAs of indicated type. MCP Class indicates the result of the sensor / kinase correlation algorithm (see section 4.2).

CheA Type			
Preliminary	Final	Phylogeny	MCP Class
ARXY	F2	Spirochetes	48H
Alpha	F5	α-proteo	38H
Ecoli	F7a, F7b	γ, β, α-proteo	36H
Ecoli99	F8	γ, δ, α-proteo	34H
Grampos	F1a, F1b	Firmicutes	44H
VAW	F3	ε-proteo	28H, 40H
WRWMAB	Alt	γ, β, δ, α-proteo	40H
YBACDR	F1a, F1b	Euryarchea	44H
YWMA	Tfp	cyano, γ, β-proteo	40H
YZAB	F6	γ-proteo	40H
uncat1		δ, γ, β-proteo	
uncat2		Firmicutes	
uncat3		Spirochetes	
uncat4	F9	δ-proteo, Firmicutes, Euryarchea	double 44H
uncat5		δ-proteo	

Table 4.2 CheA Classification. Columns as in Table 4.1, reordered by final designation from [25].

CheA Type			
Final	Preliminary	Phylogeny	MCP Class
F1a, F1b	YBACDR, Grampos	Firmicutes, Euryarchea	44H
F2	ARXY	Spirochetes	48H
F3	VAW	ε-proteo	28H, 40H
F4a, F4b		δ-proteo	40H
F5	Alpha	α-proteo	38H
F6	YZAB	γ-proteo	40H
F7a, F7b	Ecoli	γ, β, α-proteo	36H
F8	Ecoli99	γ, δ, α-proteo	34H
F9	uncat4	δ-proteo, Firmicutes, Euryarchea	double 44H
F10		δ-proteo	64H
Tfp	YWMA	cyano, γ, β-proteo	40H
Alt	WRWMAB	γ, β, δ, α-proteo	40H

4.2 Sensor / Kinase Correlation Algorithm

In organisms with multiple chemotaxis modules and many receptors scattered through the genome, it is an important goal to determine with which module each receptor in the genome interacts. That correlation can be done using an algorithm that borrows features from phylogenetic profiling [201-203] and from a technique called matrix alignment [204].

The goal of the standard phylogenetic profiling algorithm is to find sets of interacting proteins within a set of N genomes. The first step is to find the set of orthologs for each protein of interest in each genome. Orthologs are often defined methodologically as reciprocal best BLAST hits or members of the same COG [132,133]. (In our algorithm, receptor orthologs are defined as members of the same length class and kinase orthologs as members of the same CheA subfamily.) The next step is to generate an N-bit binary string for each type of protein, where the n^{th} bit is one if the n^{th} genome contains that kind of protein, and zero otherwise. This binary string is the phylogenetic profile of that set of orthologous proteins. Protein sets that have the same or similar profiles are predicted to interact [202]. Using phylogenetic profiles, entire networks of multiple interacting proteins can be reconstructed from genomic data [203].

The assumption behind matrix alignment is that when phylogenetic trees of two interacting protein families have clusters that can be aligned, then subfamilies from the aligned clusters can be predicted to interact [205]. A matrix of distances between each protein in a multiple sequence alignment forms the foundation of distance-based phylogenetic tree-building methods like neighbor-joining. In matrix alignment, the distance matrices of two interacting protein families are compared by computing the root mean square deviation (r.m.s.d.) between their elements. Columns and rows of one of the matrices are re-ordered until this difference is minimized; this re-ordering is equivalent to inverting the order of leaves in subtrees of the associated phylogenetic tree, a process that does not change the tree topology. After matrix alignment, the leaf order of the two trees

is congruent, and each pair of leaves in the same place on the two trees is predicted to interact.

What the sensor / kinase algorithm shares with the above techniques is the prediction of interaction between aligned clusters on two phylogenetic trees. The benefit of our algorithm over matrix alignment is that it allows many-to-one correlation, i.e. association of many sensors with one kinase, while matrix alignment can find only one-to-one correlations [204]. The sensor / kinase correlation algorithm, though, is limited by its reliance on the prior existence of subfamily categorizations for each protein.

The first step in the sensor / kinase correlation algorithm is to build a tree of all the MCPs. The next step is to associate with each leaf of the tree, i.e. with each MCP sequence, a 'CheA type profile' consisting of a 15-bit binary string. For each of the preliminary CheA types in Table 4.1, a bit in the string is set to one if that type of CheA is present in the genome containing the MCP; otherwise the bit is set to zero. The next step in the algorithm is to partition the MCP tree into maximal subtrees containing MCPs of only one length class. Finally, the MCPs within each subtree are predicted to interact with the CheA type associated with all of them (Figure 4.3). Very rarely, especially for small subtrees, multiple CheA types are associated with all of the MCPs in the subtree, and then no more specific prediction can be made. Similarly – and this is a more difficult problem – if multiple CheAs of the same type are present in a genome, the algorithm cannot predict with which of those CheAs the MCPs in a subtree interact.

An important variable in the sensor / kinase correlation algorithm is the method used to build the MCP tree. The method of calculating distance has a significant impact on the tree output from the NJ method. Another issue is how to deal with alignment gaps, since gaps also affect the calculation of distance between sequences. The NJ method in MEGA 3.1 has two options for dealing with gaps: complete deletion and pairwise deletion. In complete deletion, all columns in the alignment that contain gaps in any sequence are removed before calculating distances. In pairwise deletion, gaps are only

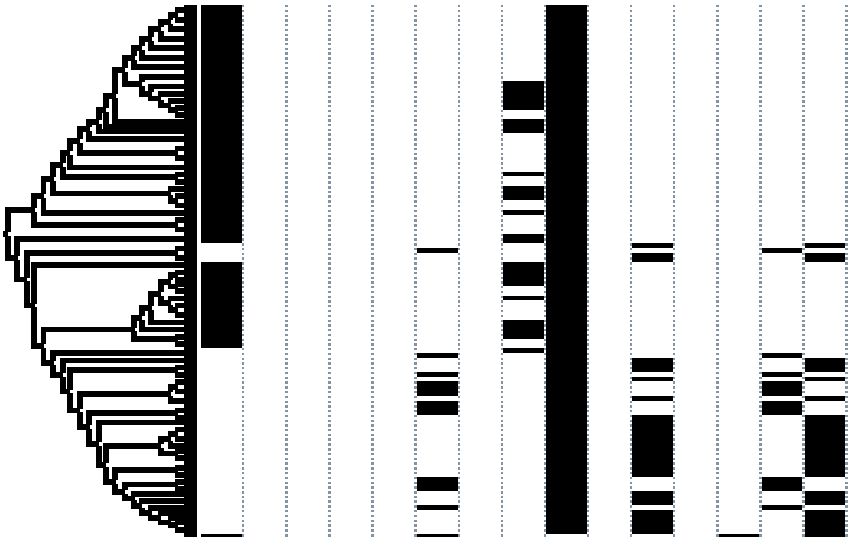

Figure 4.3 Illustration of the sensor / kinase correlation algorithm. A subtree of the full MCP tree is shown. All MCPs in the subtree are of class 38H and come from α-proteobacterial genomes. The fifteen columns at right represent the CheA type profile of each MCP. Each column represents presence or absence in the genome of one of the fifteen preliminary CheA classes from Table 4.1. Thus all MCPs in a genome share the same CheA type profile. While many of the genomes have multiple CheA types, only one type is common to all of the MCPs in this subtree, namely the F5 / Alpha type.

removed from distance calculations involving sequences that have gaps in a given column. Our initial tests of the algorithm used a simple distance calculation based on the number of amino acid differences (N_{aa}) between each sequence pair and complete deletion of columns with gaps. The N_{aa} distance calculation method, though, is not suitable for use with pairwise gap deletion [206]. Pairwise gap deletion is preferable, since it allows kinase interaction predictions to be made for class 24H MCPs and MCPs with gaps relative to their subfamily model. Including these sequences in a tree built with complete gap deletion removes too many alignment columns from consideration, so not enough information remains to build an accurate tree.

Tests using pairwise gap deletion with the Poisson correction and Equal Input models for distance estimation [206] gave similar results. Compared to the tree built from ungapped sequences with complete gap deletion N_{aa} distances, Poisson correction and

Equal Input had 19 and 13 different predictions, respectively. The predictions reported in the Cheops database are based on the results from the Equal Input model. We applied the sensor / kinase correlation algorithm to 1835 MCPs from the major classes and correlated 1727 (94%) of them unambiguously to a specific CheA type. Of these, 1522 (88% / 83% of original) were correlated to a specific CheA since only one CheA of that type was found in the genome.

Of 280 CheAs in 152 genomes, 188 (67%) had direct MCP associations, while an additional 48 had partial associations – the MCP was associated with a CheA of that type, but there were multiple CheAs of that type in the genome. In all, 236 of 280 CheAs (84%) had some level of association. There are many interesting ways to parse this data; Table 4.3 shows the number of CheAs with particular numbers of direct MCP associations, which is interesting to think about in terms of signal integration. How many sensory inputs can the molecular architecture of the chemoreceptor-kinase complex actually integrate simultaneously? Are the 46 receptors associated with the F6 / YZAB of *Vibrio vulnificus YJ016* all expressed in the array at the same time, or do different

Table 4.3 Results of the Sensor / Kinase Correlation Algorithm. For the 188 kinases with direct associations, this table lists the number of CheAs associated with specific numbers of sensors. The greatest number have just one association, but there are a significant number that interact with multiple sensors, up to a maximum of 46.

N MCPs	N CheAs with N MCPs
1	55
2	24
3	12
4	6
5	15
6-10	31
11-15	11
16-20	15
21-25	5
26+ (max 46)	14

environmental conditions trigger the expression of different subsets of receptors? The latter seems more likely, and it would be interesting to try to identify the regulatory framework controlling the expression of receptors.

Although there is certainly an error rate associated with the sensor / kinase correlation algorithm, and some interactions are wrongly predicted, the error is difficult to estimate. An improvement of the algorithm would be to use bootstrapping, running it 1000 times and measuring the accuracy of each prediction based on the number of times the correlated CheA type changes between runs. The reason that bootstrapping has not yet been implemented is technical; MEGA 3.1 runs on a Windows PC, and the algorithm that traverses the tree structure and makes the predictions runs under Linux, so there is currently a manual step involving transfer of the tree between operating systems.

4.3 Case Study: Evolution of Chemotaxis in Epsilon-Proteobacteria

We will look now at a case study, the ε-proteobacteria, and show how integrating information about receptor and kinase diversity can provide functional and evolutionary insights that are uniquely available from a comparative genomic approach. The chemotaxis module in ε-proteobacteria is classified as F3 / VAW; Figure 4.2A shows a detailed view of this module on the CheA classification tree. Note that many components of the module are not encoded in the CheA gene neighborhood. In *Campylobacter jejuni*, there is an unusual lack of operon structure for all functional modules [207]; perhaps this is a general feature of the gene regulatory mechanism in ε-proteobacteria.

Pictured in Figure 4.4 is a schematic diagram of the F3 / VAW chemotaxis architecture found in ε-proteobacteria. The system is very similar to that found in *E. coli*, except that there is no feedback from CheA to CheB via phosphorylation, because the CheB in ε-proteobacteria does not contain a receiver domain. Thinking back to the epistemological disagreement about the molecular basis of adaptive feedback in *E. coli* (section 1.5), the lack of a receiver domain on CheB in ε-proteobacteria supports the

explanation from robustness modeling that precise adaptation results from activity-dependent kinetics of CheB and does not require feedback via CheB phosphorylation. Interestingly, the CheA in all ε-proteobacteria does have a receiver domain (domain P6), so there is an autophosphorylation loop whose function has not been precisely determined. ε-proteobacteria also have a CheV scaffold protein, consisting of the CheW scaffold domain and a receiver domain. CheV may have active and inactive conformations based on its phosphorylation state, and thus may provide an alternative feedback mechanism. For a long time, no CheZ sequences were identified in ε-proteobacteria, but Kristin Wuichet has determined that the CheZ domain model from Pfam (accession PF04344) is inadequate. Using PSI-BLAST analysis, she found divergent CheZ proteins in δ-, α-, and ε-proteobacteria [25]. CheZ has also recently been identified experimentally in *Helicobacter pylori* [208].

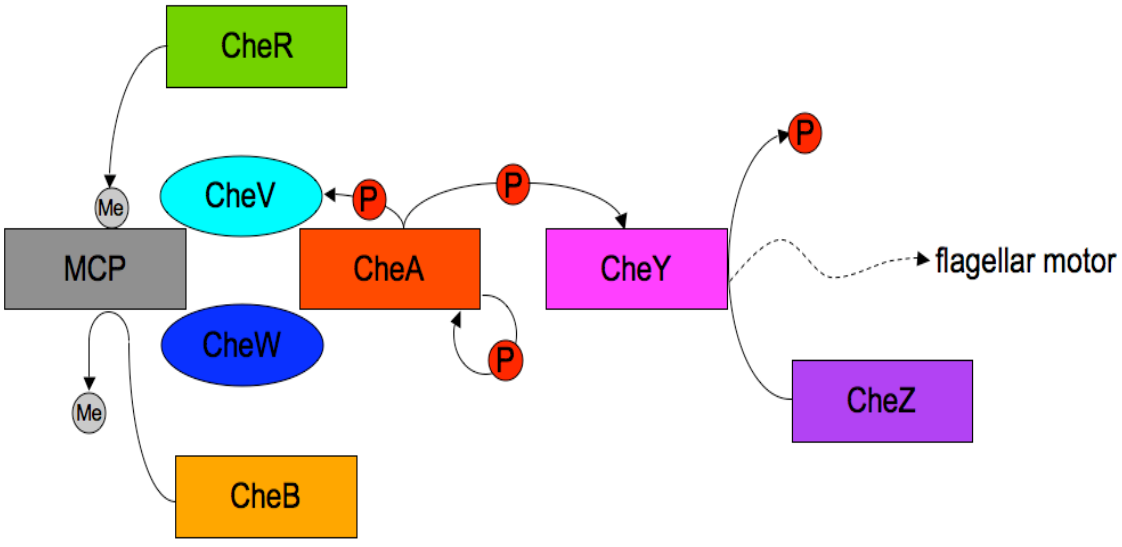

Figure 4.4 The network architecture of chemotaxis in ε-proteobacteria.

All ε-proteobacteria have only one CheA, but they all have two length classes of MCP, namely 40H and 28H. The 28H MCP class is found exclusively in ε-proteobacteria. The presence of two length classes of MCP generates several alternative hypotheses regarding the structure of chemoreceptor arrays in this phlyogenetic group. One possibility is that both 40H and 28H MCPs are present simultaneously in the same chemoreceptor array. This possibility seems unlikely, because it would require a significant reorganization of the interactions between receptor, scaffold, and kinase over a fairly short evolutionary timespan. The other option is that the 40H and 28H MCPs are present in distinct chemoreceptor arrays that are expressed separately either in space or in time. Spatial separation of multiple chemoreceptor arrays should be easy to check experimentally by expressing fluorescent hybrids of both MCP classes to look for co-occurrence [194]. We prefer the latter hypothesis, that ε-proteobacteria express different MCPs at different phases of their life cycle. The interesting question is to determine what the different MCP classes sense, and to hypothesize about when in the life cycle the different arrays are utilized.

The 40H receptor in *Helicobacter pylori*, tlpB, has been found to control negative taxis away from regions of high acidity [209]. *H. pylori* is the causative agent of gastric ulcers in humans [210]; an important stage in its life cycle is navigating towards the gastric mucosa, which is a more neutral pH than the rest of the stomach. We propose that a tlpB-containing chemoreceptor array assembles during this stage of the life cycle, and that the other 3 MCPs in *H. pylori*, because of their different length, are expressed during some earlier or later phase (Figure 4.5).

Another interesting fact relates to the methylation sites on 28H MCPs. Shown in Figure 4.6 are the two adjacent sites of methylation, in heptads N15 and N14, that are conserved in 28H MCPs. A 28H MCP dimer actually then has four sites of methylation. *H. pylori* is unique among ε-proteobacteria in having lost its CheR and CheB adaptation

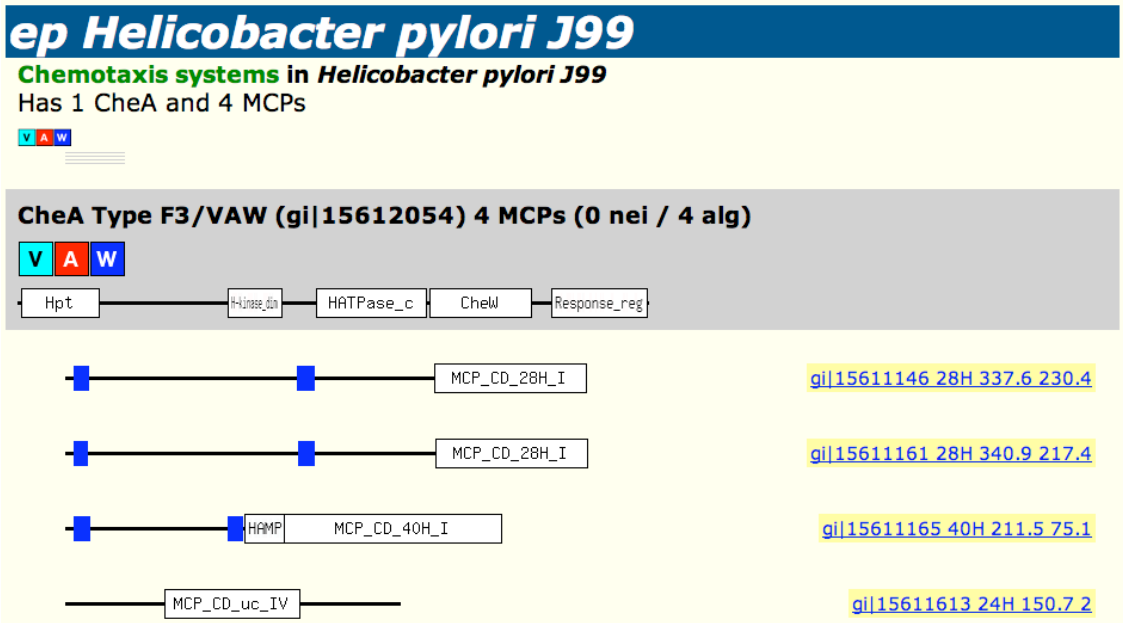

Figure 4.5 Detail view from the Cheops database of chemotaxis pathways in *Helicobacter pylori J99*.

Figure 4.6 Characteristic methylation sites in 28H MCPs. (A) Sequence logo showing the conserved methylation sites in heptads N15 and N14 of 28H MCPs. (B) Heptad register schematic showing how methylation sites map onto the surface of the four helical bundle structure of the MCP dimer.

Table 4.4 Loss of N15 and N14 methylation sites in the two 28H MCPs from *H. pylori*. The Genbank ID refers to strain J99, but the sequences are identical in strain 26695. The sequence shown is from alignment position N15b to N14g.

Genbank ID	Sequence	Meth site?
15611146	KNTTQS L EEITNI	--
15611161	METSKT I ENITTS	--

enzymes, and the 28H MCPs in *H. pylori* have lost the associated methylation sites (Table 4.4). Here is a set piece example of evolution at work. After losing the adaptation enzymes, selective pressure to maintain their sites of interaction with the MCP was lost, and so the conserved methylation sites degraded.

Referring back to our earlier discussion of evolvability in signaling networks (section 1.7.2), we argue that as a pathogen, *H. pylori* is under selective pressure to reduce the number of proteins in its genome. *H. pylori* actually has three CheV proteins [211], which, with the autofeedback via the CheA receiver domain, may provide the reservoir of robustness that allows *H. pylori* to maintain some mechanism of adaptation, and the ability to perform chemotaxis, even after losing the CheR/CheB adaptation pathway. It should be possible to build a model of how chemotaxis functions in ε-proteobacteria [58,65,212], and then modify it to explain not only how chemotaxis works in *H. pylori*, but how it evolved from its ancestral state.

4.4 Single-Input Architectures and the Origin of Chemotaxis

An important insight from Kristin Wuichet's work is that chemotaxis modules can be divided into three main classes: F, Tfp, and Alt [25]. The module controlling flagellar motility (F) has diversified widely over evolutionary time into ten distinct subclasses (Table 4.2). Another module (Tfp) controls twitching motility via Type-IV pili [15,213], and a third module (Alt) has been co-opted to control alternate outputs unrelated to motility [22]. Table 4.5 and Table 4.6 show an interesting feature of the Tfp and Alt systems, namely that each Tfp or Alt module associates with only one MCP. Both systems interact with an MCP of subfamily 40H, which is usually located in the CheA gene neighborhood.

There are exceptions in both tables, but they do not invalidate the trend. In each case, the data can be explained by mis-predictions by the algorithm, draft status of the genome, or specific evolutionary conditions experienced by the organism. For example,

members of the *Pseudomonas* group exhibit both flagellar motility controlled by an F6 / YZAB chemotaxis system [194] and pili-based motility controlled by a Tfp / YWMA chemotaxis system [15,213]. Both the F6 and Tfp systems utilize class 40H MCPs. *P. aeruginosa* has 18 40H MCPs associated with its F6 system and 2 40H MCPs associated with its Tfp system (data available in Cheops database; see chapter 5). This is probably a case of mis-prediction by the sensor / kinase algorithm because of the presence of so many 40H MCPs in the genome. *Deinococcus radiodurans* is a unique organism that is highly resistant to dessication and ionizing radiation [214]. It is thought not to be motile, so its residual Tfp-class CheA probably plays some role uniquely evolved in that organism. The three 40H MCPs associated with this CheA are collinear within the CheA neighborhood and appear to result from a recent duplication.

Exceptions aside, the overall trend of the data implies that the Tfp and Alt chemotaxis modules are single-input / single-output systems. Apparently Tfp and Alt chemoreceptor arrays do not perform a signal integration function, although MCP clustering within the array might still play a role in signal amplification [187,188]. It is clear from Figure 4.2B that there are two Tfp / YWMA subclasses. In the proteobacterial group, there has been a proliferation of phosphotransfer (Hpt / P1) domains in the kinase, which probably generate complicated signaling and feedback mechanisms that compensate for the lack of signal integration through the receptors [213]. In the cyanobacterial group, there has been a proliferation of entire chemotaxis modules (Table 4.5). Cyanobacteria respond to multiple sensory inputs by encoding multiple chemotaxis systems, one controlling each input [87].

Table 4.5 Tfp / YWMA is a single-input module. In most organisms, each kinase of type Tfp / YWMA associates with only one MCP. See Table 4.6 for column definitions.

Tax	Organism	Draft?	N MCP	N CheA	Diff.
cy	Crocosphaera watsonii WH 8501	Y	2	4	-2
cy	Anabaena variabilis ATCC 29413	Y	3	3	0
cy	Nostoc sp	N	3	3	0
cy	Synechococcus elongatus PCC 6301	N	2	2	0
cy	Synechococcus elongatus PCC 7942	Y	2	2	0
cy	Synechocystis PCC6803	N	3	3	0
cy	Thermosynechococcus elongatus	N	3	3	0
cy	Trichodesmium erythraeum IMS101	Y	2	2	0
bp	Azoarcus sp EbN1	N	1	1	0
bp	Methylobacillus flagellatus KT	Y	1	1	0
bp	Nitrosomonas europaea	N	1	1	0
bp	Polaromonas sp. JS666	Y	1	1	0
gp	Pseudomonas putida KT2440	N	1	1	0
gp	Pseudomonas syringae	N	1	1	0
gp	Pseudomonas syringae pv B728a	N	1	1	0
gp	Psychrobacter sp. 273-4	Y	1	1	0
bp	Ralstonia eutropha JMP134	Y	1	1	0
bp	Ralstonia metallidurans CH34	Y	1	1	0
bp	Ralstonia solanacearum	N	1	1	0
bp	Rubrivivax gelatinosus PM1	Y	1	1	0
gp	Saccharophagus degradans 2-40	Y	2	2	0
bp	Thiobacillus denitrificans ATCC 25259	Y	1	1	0
gp	Xanthomonas campestris	N	1	1	0
gp	Xanthomonas citri	N	1	1	0
gp	Xanthomonas oryzae KACC10331	N	1	1	0
gp	Xylella fastidiosa	N	1	1	0
gp	Xylella fastidiosa Ann-1	Y	1	1	0
gp	Xylella fastidiosa Dixon	Y	1	1	0
gp	Xylella fastidiosa Temecula1	N	1	1	0
gp	Pseudomonas aeruginosa	N	2	1	1
gp	Pseudomonas aeruginosa UCBPP-PA14	Y	2	1	1
gp	Pseudomonas fluorescens PfO-1	Y	2	1	1
gp	Xanthomonas campestris 8004	N	1	0	1
de	Deinococcus radiodurans	N	3	1	2
cy	Nostoc punctiforme PCC 73102	Y	6	4	2
gp	Pseudomonas fluorescens Pf-5	N	3	1	2
bp	Chromobacterium violaceum	N	3	0	3
bp	Dechloromonas aromatica RCB	Y	5	1	4

Table 4.6 Alt / WRWMAB is a single-input module. In most organisms, each kinase of type Alt / WRWMAB associates with only one MCP. Column definitions: N MCP, Number of MCPs in the genome associated with this type of CheA by the sensor / kinase correlation algorithm. N CheA, Number of CheAs of this type found in the genome. Diff, N MCP - N CheA. If Diff = 0, the single-input prediction holds, since there is a one-to-one correspondence between MCP and CheA. Tax, taxonomy: cy, cyanobacteria; ap, bp, gp, dp, mp: α-, β-, γ-, δ-, magneto-proteobacteria, respectively; ch, chloroflexi; de, Deinococcus. Draft, draft vs. complete genome sequence.

Tax	Organism	Draft?	N MCP	N CheA	Diff.
mp	Magnetococcus sp. MC-1	Y	1	2	-1
ap	Mesorhizobium sp. BNC1	Y	1	1	0
bp	Burkholderia cepacia R1808	Y	1	1	0
bp	Burkholderia cepacia R18194	Y	1	1	0
bp	Ralstonia eutropha JMP134	Y	1	1	0
bp	Ralstonia metallidurans CH34	Y	1	1	0
ch	Chloroflexus aurantiacus	Y	1	1	0
dp	Geobacter metallireducens GS-15	Y	1	1	0
gp	Pseudomonas aeruginosa UCBPP-PA14	Y	1	1	0
gp	Pseudomonas fluorescens PfO-1	Y	1	1	0
ap	Mesorhizobium loti	N	1	1	0
ap	Sinorhizobium meliloti	N	1	1	0
bp	Burkholderia mallei ATCC 23344	N	1	1	0
bp	Burkholderia pseudomallei K96243	N	1	1	0
gp	Pseudomonas aeruginosa	N	1	1	0
gp	Pseudomonas fluorescens Pf-5	N	1	1	0
gp	Pseudomonas putida KT2440	N	1	1	0
gp	Pseudomonas syringae	N	1	1	0
gp	Pseudomonas syringae pv B728a	N	1	1	0
gp	Methylococcus capsulatus Bath	N	2	1	1
bp	Polaromonas sp. JS666	Y	3	1	2

The fact that Tfp / YWMA and Alt / WRWMAB are single-input modules raises interesting issues regarding the evolutionary history of chemotaxis. Recall from section 1.2 the hypothesis that chemotaxis originated when a Class I histidine kinase split into separate sensor and kinase proteins, and that natural selection favored this innovation because it led to the chemoreceptor array structure with its ability to integrate multiple sensory inputs. If chemotaxis did evolve as a signal integration module, then the single-input Tfp and Alt systems are degenerate and have lost the ability. The alternative is that the Tfp and Alt systems represent a residual early phase of chemotaxis evolution and that the ability to integrate signals evolved later.

A related issue is the evolutionary history of the MCP cytoplasmic domain. The two oldest MCP_CD subfamilies are 44H, present in Firmicutes and Archea, and 40H, the most abundant class, present in cyanobacteria and proteobacteria. 44H MCPs are associated only with F1 and the rare F9 CheA types, while 40H MCPs are associated with F3, F4, F6, Tfp, and Alt CheA types (Table 4.2). The consensus in the field of molecular phylogenetics is that Firmicutes are older than Cyanobacteria [215,216], although a minority holds the opposite view [217].

These three evolutionary questions – whether chemotaxis originated as a signal integration module, which length class of MCPs is oldest, and which clade of bacteria are oldest – are all related. If Firmicutes are older than cyanobacteria, then chemotaxis originated to control flagellar motility with an integrated response to multiple sensory inputs mediated by 44H MCPs. If cyanobacteria are older than Firmicutes, then chemotaxis originated to control pili-based motility in response to a single sensory input mediated by 40H MCPs. The two Tfp / YWMA subfamilies established alternate routes to increasing the complexity of their signaling mechanism, while in this scenario the ability to integrate sensory inputs developed later in the F system. The fact that all five minor classes of MCPs evolved by a pair of symmetric insertions relative to their parental class shows that both insertions and deletions are possible in the domain (see section 3.4),

so 40H could be the oldest class. Nevertheless, in agreement with the phylogenetic consensus, we argue that 44H MCPs are the oldest class and present a possible evolutionary scenario.

One of the methylation sites in the 44H receptors of *B. subtilis* is at alignment position C18c; 40H MCPs have a gap at that location (Figure 3.2B, Figure A.1). If 40H MCPs evolved from 44H MCPs, then their birth may have been marked by the loss of this key methylation site, which could have disrupted the methylation-dependent adaptation mechanism in systems that use 40H MCPs. In fact, most 40H MCPs lack methylation sites that match the global consensus motif (Figure 3.6). Perhaps disruption of methylation-dependent adaptation led the first 40H-containing organism to fall back on a methylation-independent adaptation mechanism that was incapable of signal integration in the chemoreceptor array. Methylation-based adaptation might have been restored later when class 36H was derived from class 40H by a further deletion in the FBS. This evolutionary scenario provides a plausible explanation for why the adaptation mechanisms of *E. coli* and *B. subtilis* differ (see section 1.8.6), but it does not clearly explain the presence of CheB and CheR in F6 / YZAB and some Tfp / YWMA and Alt / WRWMAB systems. To resolve these issues, more experimental work needs to be done on the adaptation mechanism of systems that use 40H MCPs. Although most 40H MCPs lack methylation sites matching the global consensus motif, the F6 / YZAB system in particular may have a divergent motif containing a conserved large valine residue in the g heptad register where the global consensus has a small residue (Figure 3.6). This methylation site may have co-evolved with the active site of the associated adaptation enzymes.

The discussion above should make clear the integrative power of comparative genomics. With just a few more data points from one of several divergent fields of study, it may soon be possible to resolve the early history of chemotaxis more definitively.

CHAPTER 5

DEVELOPMENT OF A CHEMOTAXIS DATABASE

The Cheops (Chemotaxis operons) database, available at http://genomics.ornl.gov/cheops/, has been a central component of this research project, enabling me to visualize the chemotaxis proteins and pathways present in any given genome, to explore the diversity of chemoreceptor classes, and to explore the results of the sensor / kinase correlation algorithm. The goal for Cheops development is to integrate it into the MiST database [125], since that would make MiST a comprehensive resource for exploring prokaryotic signal transduction. MiST currently focuses on one- and two-component systems [24]; chemotaxis proteins are included, but they are not organized into pathways or analyzed in a systematic way. To integrate Cheops into MiST it is necessary to automate the process of chemotaxis pathway deduction for newly sequenced genomes. The two central facets of automatic pathway prediction are determination of receptor and kinase subfamilies. Automatic determination of MCP class has been finished and an MCP prediction server is available on Cheops. Automatic determination of CheA type is not yet complete, and progress towards that goal will be described in the next chapter.

5.1 Cheops Database

The Cheops database is built on a developmental version of the MiST database [125,126] that contains information from 236 complete and 76 draft genomes (Figure A.1). Of those, 152 genomes (106 complete and 46 draft genomes) contain chemotaxis systems. For every organism in the database two useful visualizations are available. The overview or genome picture shows snapshots of each chromosome with the chemotaxis proteins displayed in their chromosomal locations (Figure 5.1). Because draft genomes

contain many chromosomal fragments, the overview is currently disabled for draft genomes. The detailed view displays information on each chemotaxis pathway, showing CheA gene neighborhood and domain structure (Figure 5.2). In genomes with multiple chemotaxis pathways, the detailed view utilizes the sensor / kinase correlation algorithm (see section 4.2) to sort the pathways based on the number of associated MCPs, since the pathway with the most associated MCPs is usually the one responsible for chemotaxis. Listed under each pathway are domain models of the MCPs associated with it by the algorithm. Domain visualization utilizes the archviz.pl Perl script developed by Luke Ulrich for the MiST database [125,126]. MCPs that the algorithm was unable to associate with a specific CheA are listed separately at the end. The MCP domain models show information about sensory membrane topology (section 2.12) and cytoplasmic domain subfamily (section 2.9).

In the early phases of this project, the overview was useful because it highlighted the fact that most MCPs are scattered around the genome outside of CheA neighborhoods, while most other chemotaxis proteins are located near CheA. The detailed view was central to the analysis of receptor diversity and evolution [37] and is a useful tool for experimentalists, especially those working in organisms with multiple pathways.

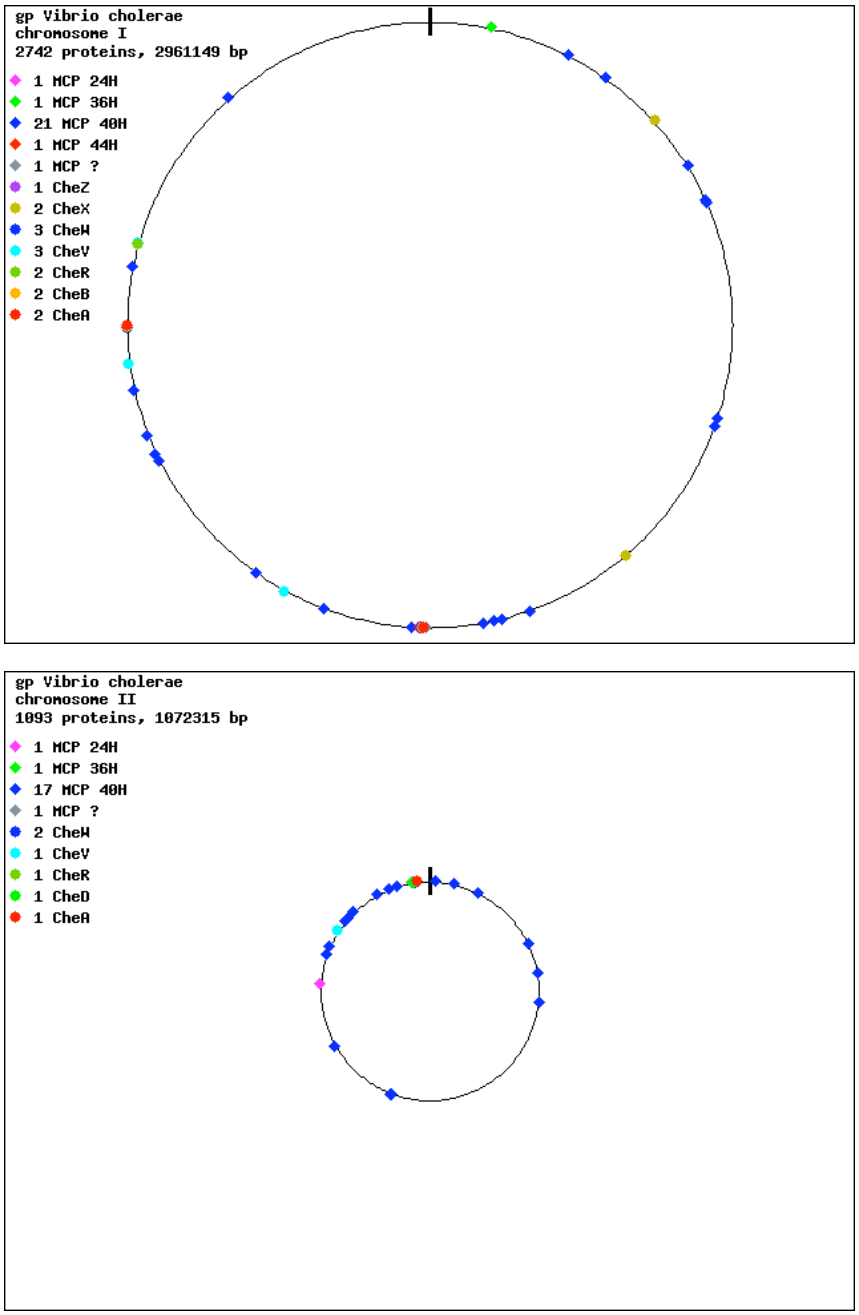

Figure 5.1 Overview of chemotaxis proteins in *Vibrio cholerae* from the Cheops database. The distribution of chemotaxis proteins on each of *V. cholerae*'s two chromosomes is visualized separately. Chromosomes and plasmids are listed in order of decreasing size scaled to the largest component in each genome; figures are not to scale between genomes. Linear chromosomes are pictured as circles with 5% of missing arc. Position zero, usually near the origin of replication, is indicated by a bar at top. MCPs are indicated by diamonds; other chemotaxis proteins by circles. MCP cytoplasmic domain subfamilies are uniquely colored to differentiate them in the image, while the color of other chemotaxis proteins matches that in the detailed view.

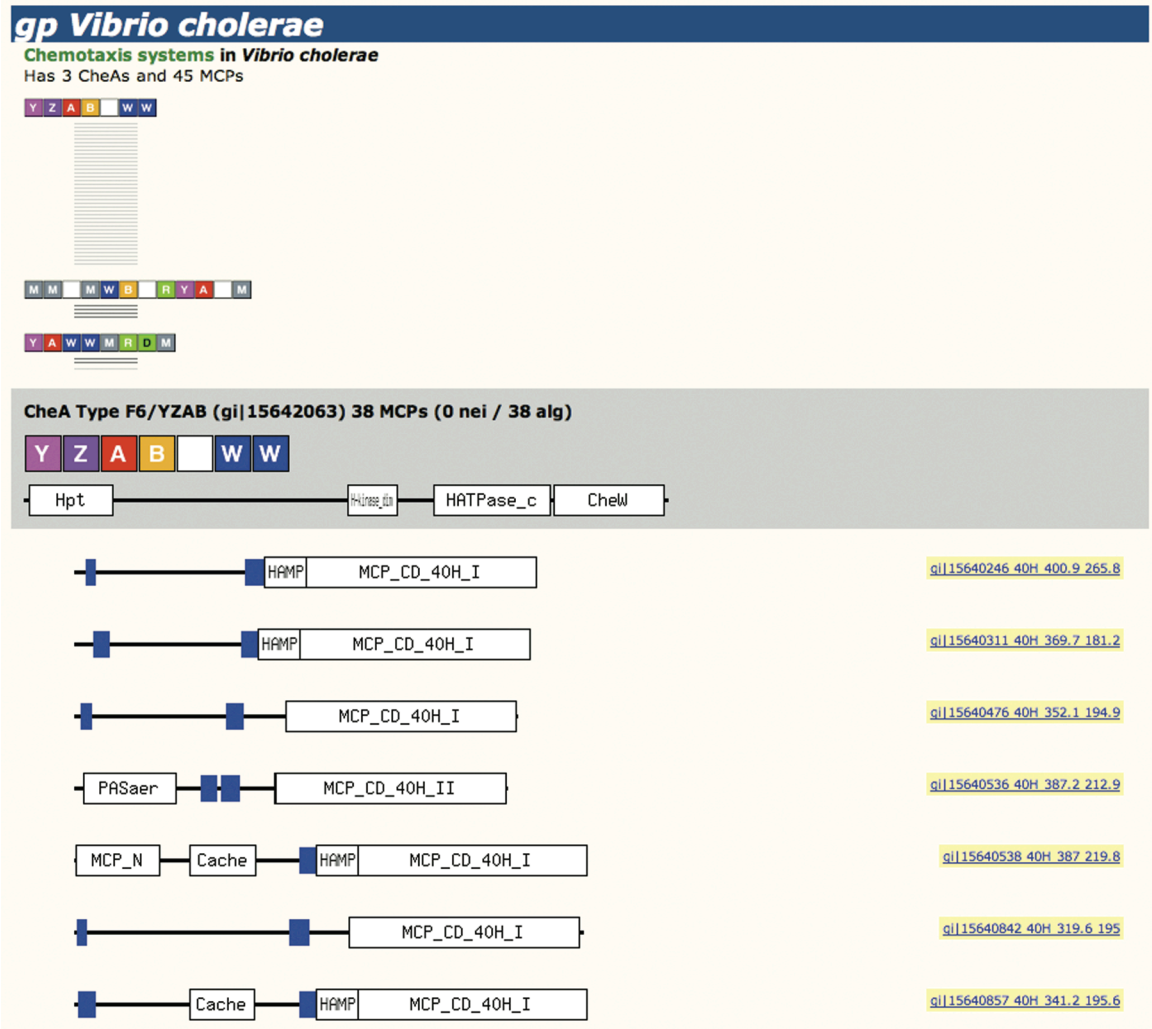

Figure 5.2 Detailed view from the Cheops database of chemotaxis pathways in *Vibrio cholerae*. At the top is a reduced representation of all CheA neighborhoods and the MCPs associated with them by the sensor / kinase correlation algorithm. MCPs in the neighborhood of CheA are dark lines, while those correlated to the CheA only by the algorithm are grey lines. Clicking on a pathway icon jumps to that pathway in the listing. Below this overview, each CheA is listed in descending order of the number of MCPs associated with it. Detailed information about each CheA is in grey shading. The first line shows CheA type as outlined in section 4.1, then Genbank ID and the number of MCP associations broken down by neighborhood vs. algorithm. The next line shows the CheA gene neighborhood. Below that is the CheA domain organization. Below the grey shaded Chea information are listed all MCPs associated with that CheA, pictured by their domain organization. Transmembrane regions are in blue. C-terminal pentapeptide tethers are indicated in red. Within the MCP cytoplasmic domain both MCP_CD subfamily and sensor class are identified. At right is the MCP's Genbank number, highest-scoring subfamily, and the scores of the top two subfamilies. Clicking on this text calls up a screen showing detailed information about the MCP_CD subfamily prediction which is useful for manually determining subfamily for sequences where automatic prediction fails (see section 5.2).

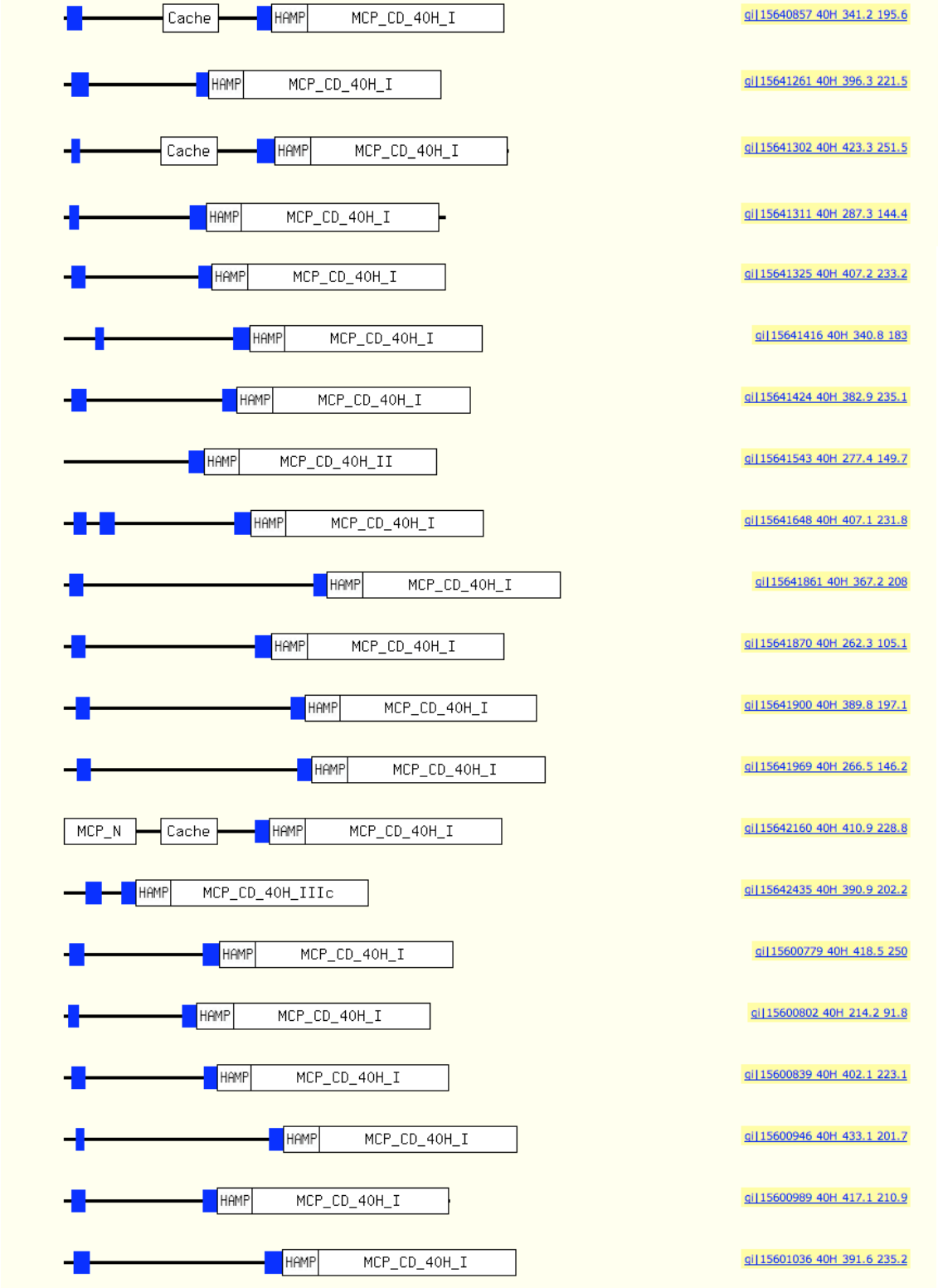

Figure 5.2 continued.

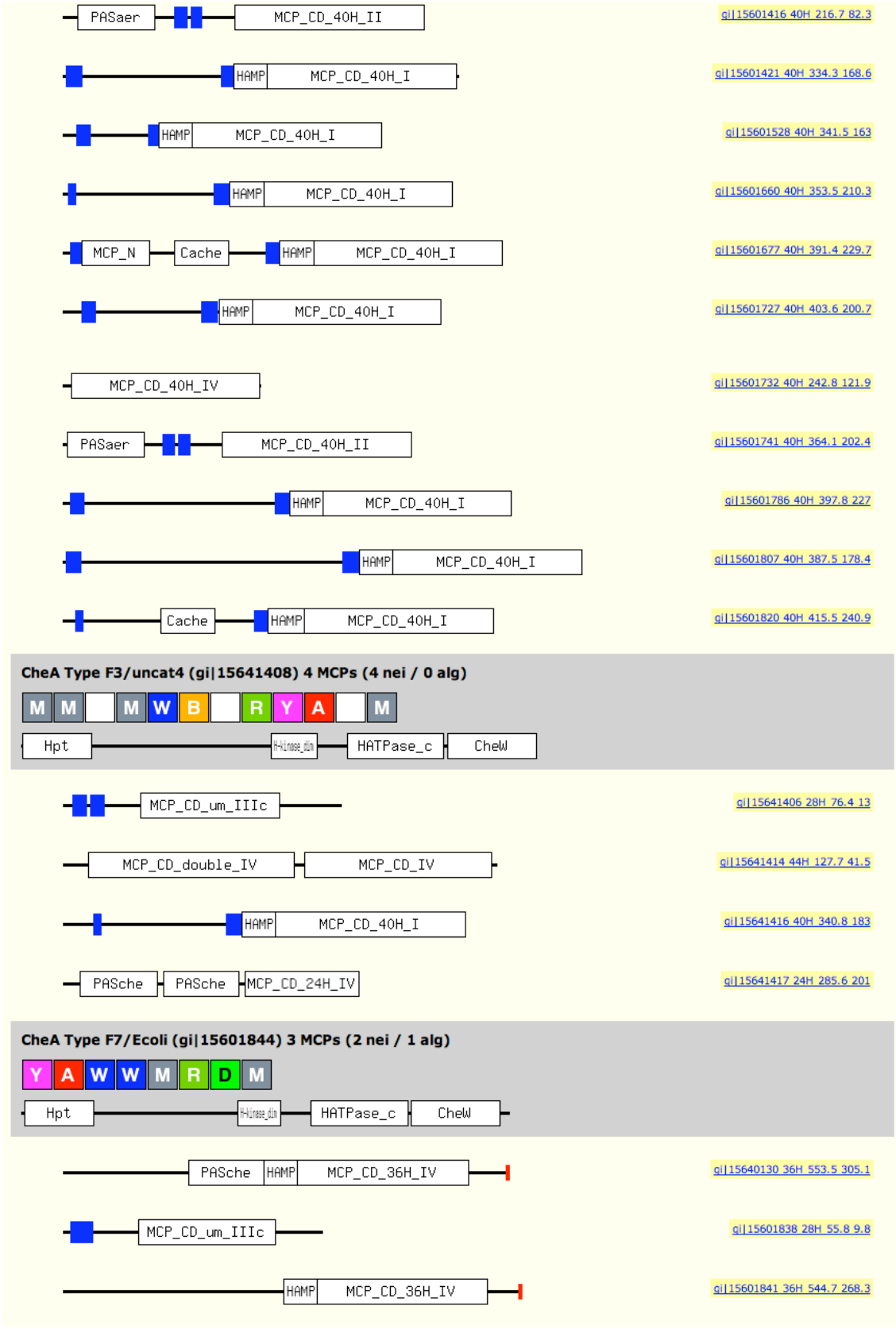

Figure 5.2 continued.

5.2 MCP Prediction Server

An important aspect of analyzing newly sequenced genomes is predicting the subfamily to which the cytoplasmic domains of MCPs belong. There is an MCP prediction server available at the Cheops website to facilitate this analysis. This tool should be useful both for casual analysis and for systematic analysis of MCPs from newly sequenced genomes while automatic determination of CheA type is still under development.

For example, a quick perusal of the list of completely sequenced genomes at NCBI (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/Complete.txt) shows that a new Geobacter genome, *Geobacter uraniumreducens Rf4*, became available in May 2007. It might be interesting to examine the MCPs in this newly available genome, since Geobacter species are among those delta-proteobacteria that have very long 40+24H (64H) MCPs. All the MCPs in this species can be quickly downloaded from the NCBI website (http://www.ncbi.nlm.nih.gov/) by searching the Entrez Protein database with the query ‘“"Geobacter uraniumreducens Rf4"[Organism] AND "methyl-accepting"[Title] AND refseq[filter]”. This query takes advantage of the fact that sequences that match the Pfam MCPsignal domain model are annotated with the title ‘methyl-accepting chemotaxis protein.’ Without the Refseq filter, each of the 27 MCPs in this organism is listed twice because of the redundancy in Genbank. The Refseq entry refers to the final, official protein from the complete genome.

Figure 5.3 shows the result of downloading these 27 MCP sequences from Entrez Protein in fasta text format, pasting them into the MCP prediction server query box, and performing the analysis. We see that *G. uraniumreducens Rf4* has a diversity of MCP classes, including 4 64H MCPs, almost doubling the number of 64H MCPs known when they were first discovered [37]. There are 5 MCPs in the genome that are not well-predicted because the difference in score between the top two matching classes is less than 50 units (see section 2.9); this parameter is adjustable on the prediction server.

Clicking on any sequence name in the results table brings up a screen showing the details of the HMMer output, sorted in order of score. Doing this for sequence 2 reveals that it is an additional 64H MCP since the 64H model has the highest score and matches without gaps.

It is important to note that the MCP prediction server is limited to finding MCPs of the 12 classes that have already been defined. We fully expect new length classes to appear as the number of sequenced genomes increases. Because of the unique structure of the MCP dimer coiled coil, we expect most new classes to exhibit pairs of heptad-length indels located in the N- and C-terminal helical arms equidistant from the hairpin turn. In order to find these new classes, it will be necessary periodically to start from scratch by going through the whole alignment process outlined in section 2.9. This procedure is robust and will result in the definition of HMMs for the larger set of length classes, including perhaps redefinition of the HMMs for existing classes based on a larger sample size.

	Sequence name	Prediction	1st	2nd	Score	Diff
1	gi|148266411|ref|YP_001233117.1| methyl-accepting chemotaxis sensory transducer [Geobacter uraniumreducens Rf4]	40H	40H	24H	332.3	140.2
2	gi|148266232|ref|YP_001232938.1| methyl-accepting chemotaxis sensory transducer [Geobacter uraniumreducens Rf4]	---	64H	44H	171.0	28.4
3	gi|148266180|ref|YP_001232886.1| methyl-accepting chemotaxis sensory transducer [Geobacter uraniumreducens Rf4]	40H	40H	44H	233.4	61.9
4	gi|148266029|ref|YP_001232735.1| methyl-accepting chemotaxis sensory transducer [Geobacter uraniumreducens Rf4]	64H	64H	40H	536.9	392.9
5	gi|148265417|ref|YP_001232123.1| methyl-accepting chemotaxis sensory transducer [Geobacter uraniumreducens Rf4]	40H	40H	44H	310.9	76.9
6	gi|148265335|ref|YP_001232041.1| methyl-accepting chemotaxis sensory transducer [Geobacter uraniumreducens Rf4]	40H	40H	44H	438.2	168.2
7	gi|148265307|ref|YP_001232013.1| methyl-accepting chemotaxis sensory transducer [Geobacter uraniumreducens Rf4]	64H	64H	44H	930.3	708.6
8	gi|148265101|ref|YP_001231807.1| methyl-accepting chemotaxis sensory transducer [Geobacter uraniumreducens Rf4]	40H	40H	44H	398.0	143
9	gi|148265084|ref|YP_001231790.1| methyl-accepting chemotaxis sensory transducer [Geobacter uraniumreducens Rf4]	40H	40H	44H	406.5	140.6
10	gi|148265035|ref|YP_001231741.1| methyl-accepting chemotaxis sensory transducer [Geobacter uraniumreducens Rf4]	---	28H	24H	4.4	103.6
11	gi|148265031|ref|YP_001231737.1| methyl-accepting chemotaxis sensory transducer [Geobacter uraniumreducens Rf4]	34H	34H	36H	552.9	273.6
12	gi|148265028|ref|YP_001231734.1| methyl-accepting chemotaxis sensory transducer [Geobacter uraniumreducens Rf4]	34H	34H	36H	550.4	286.1
13	gi|148265024|ref|YP_001231730.1| methyl-accepting chemotaxis sensory transducer [Geobacter uraniumreducens Rf4]	34H	34H	36H	561.8	280.4
14	gi|148264885|ref|YP_001231591.1| methyl-accepting chemotaxis sensory transducer [Geobacter uraniumreducens Rf4]	40H	40H	44H	446.6	181.5
15	gi|148264820|ref|YP_001231526.1| methyl-accepting chemotaxis sensory transducer [Geobacter uraniumreducens Rf4]	34H	34H	36H	554.5	272.7
16	gi|148264733|ref|YP_001231439.1| methyl-accepting chemotaxis sensory transducer [Geobacter uraniumreducens Rf4]	---	40H	44H	230.3	0.400000000000006
17	gi|148264528|ref|YP_001231234.1| methyl-accepting chemotaxis sensory transducer [Geobacter uraniumreducens Rf4]	---	44H	40H	115.1	28.8
18	gi|148264221|ref|YP_001230927.1| methyl-accepting chemotaxis sensory transducer [Geobacter uraniumreducens Rf4]	36H	36H	34H	617.1	322
19	gi|148264134|ref|YP_001230840.1| methyl-accepting chemotaxis sensory transducer [Geobacter uraniumreducens Rf4]	---	44H	40H	166.0	24.9
20	gi|148263262|ref|YP_001229968.1| methyl-accepting chemotaxis sensory transducer [Geobacter uraniumreducens Rf4]	34H	34H	36H	518.0	270.3
21	gi|148263033|ref|YP_001229739.1| methyl-accepting chemotaxis sensory transducer [Geobacter uraniumreducens Rf4]	40H	40H	44H	376.0	130.3
22	gi|148262801|ref|YP_001229507.1| methyl-accepting chemotaxis sensory transducer [Geobacter uraniumreducens Rf4]	44H	44H	40H	341.7	174.4
23	gi|148262691|ref|YP_001229397.1| methyl-accepting chemotaxis sensory transducer [Geobacter uraniumreducens Rf4]	40H	40H	44H	434.9	153.7
24	gi|148262558|ref|YP_001229264.1| methyl-accepting chemotaxis sensory transducer [Geobacter uraniumreducens Rf4]	40H	40H	44H	394.8	117
25	gi|148262458|ref|YP_001229164.1| methyl-accepting chemotaxis sensory transducer [Geobacter uraniumreducens Rf4]	40H	40H	44H	380.3	134.1
26	gi|148262432|ref|YP_001229138.1| methyl-accepting chemotaxis sensory transducer [Geobacter uraniumreducens Rf4]	64H	64H	40H	767.1	563.9
27	gi|148262248|ref|YP_001228954.1| methyl-accepting chemotaxis sensory transducer [Geobacter uraniumreducens Rf4]	64H	64H	40H	892.3	684.4

Figure 5.3. MCPs from *Geobacter uraniumreducens Rf4* run through the MCP prediction server. Columns '1st' and '2nd' show the highest and second-highest scoring domain models. The Score column shows the highest bit score, and the Diff column shows the difference between the two highest scores. Clicking on the sequence name opens a window showing the details of the HMMer analysis, with alignments to each of the 12 MCP_CD HMMs presented in descending order of score for that sequence.

CHAPTER 6

FUTURE WORK

6.1 Automated Analysis of Chemotaxis Pathways

MiST currently provides comprehensive analysis of one- and two-component systems in prokaryotic genomes [24,125]. Adding Cheops to MiST will add the third important prokaryotic signal transduction mechanism, chemotaxis, and thus make MiST a more complete resource for understanding all facets of microbial signal transduction. A key step in integrating Cheops into MiST is to build an automated algorithm for the analysis of chemotaxis pathways in newly sequenced genomes. The main algorithmic problem that remains unsolved is the determination of CheA type for new CheA sequences.

There are two possible approaches to automatically determining CheA type. The first approach is to build HMMs for each CheA type from subalignments of the core P3-P5 domains and establish a threshold score to differentiate subfamilies from each other. This method mirrors that currently used for MCP subfamily identification, where automatic predictions are made when the score of one HMM is at least 50 bits higher than the scores from other subfamily HMMs. The second approach relies on automating the original procedure used to delineate CheA type. As outlined in section 4.1, the preliminary method depended on partitioning a maximum likelihood CheA tree into subtrees each containing kinases with a common set of shared traits depending on phylogeny, CheA domain organization and CheA gene neighborhood. It is not feasible to recalculate the entire ML tree for each new CheA sequence, since it takes at least a day on a desktop computer [25,159]. It would be useful to test the accuracy of assuming that the new CheA belongs to the subtree of the sequence to which it is most similar, using

different similarity measures from BLAST and multiple sequence alignment. I am not aware of any methods to add leaves to a pre-existing phylogenetic tree without recalculation of the whole tree. An alternative is to use a faster tree-building method like neighbor joining to recompute the CheA tree each time a new sequence needs to be categorized.

In the sensor / kinase correlation algorithm, the MCP tree was automatically partitioned into subtrees where all sequences in a subtree were members of the same MCP_CD subfamily. So far partitioning the CheA tree into subtrees has involved expert curation, since the combination of phylogeny, CheA domain organization and CheA gene neighborhood that determines CheA type is unique to each case. The point for further research is to determine whether a logical set of rules can be found that prioritizes these three CheA traits in a way that allows tree partition. Then the CheA tree could be built using a fast method, CheA type determined automatically, and the process repeated by bootstrapping to generate a quantitative measure of the accuracy of each prediction. High-confidence predictions would be those that remained the same at each tree-building iteration.

6.2 Methylation Site Patterns in Cheops Detail View

Patterns of methylation sites in groups of interacting receptors may be an important feature of the signal integration mechanism in the chemoreceptor array. It would be useful to visualize those patterns in the same style as Figure 3.8 in the set of MCPs associated with each chemotaxis pathway by the sensor / kinase correlation algorithm. Generating such a set of images is feasible using Perl, Modeller, and Pymol scripts, and would be a useful addition to the detail view of each chemotaxis pathway in the Cheops database.

6.3 Database Integration

With Cheops and MiST, we face on a small scale what systems biologists everywhere face on a large scale, namely, the complicated issue of database integration. The main feature of MiST that is useful to apply to Cheops is its computationally intensive identification of Pfam and SMART domains in all sequenced genomes. The data in Cheops that need to be integrated into MiST are curated classifications of chemoreceptors and kinases that together determine chemotaxis network architecture. The problem is that the means of identifiying common sequences between the two databases is a moving target. The internal MiST ID numbers used in Cheops grew out of sync with the ID numbers in the new version of MiST when a server move prompted a recalculation and re-indexing of the whole database.

A useful strategy for syncing two databases that rely on sequence information but that have incompatible unique identifiers is to generate a table where each sequence is identified by a cyclic redundancy check (CRC). Cyclic redundancy checking is a technique used to ensure that a datastream has not been compromised after transit over noisy communication channels, but for our purpose it has the added benefit of providing a unique identifier to each sequence based on its contents. This procedure has the advantage of consolidating duplicate sequences in closely related genomes and thus eliminating wasteful duplicate calculations when scanning the domain model databases and performing other large-scale computations. The group responsible for the Swissprot database has a Perl toolkit called Swissknife that includes an implementation of CRC64 (http://swissknife.sourceforge.net).

6.4 Further Computational Analysis of MCPs

There are two projects that are natural to pursue as follow-up to this research. First is a campaign aimed at the systematic discovery of new sensory domains in MCP sequences. It has already been possible to categorize most MCPs into sensor classes

based on their membrane topology (see section 2.12). That membrane topology information is the first step in a project to PSI-BLAST sensory regions of MCPs in the search for new sensor domain families. The major undeveloped piece of the project is automatic selection of appropriate sequences for each new iteration of PSI-BLAST so that false positive hits do not drown out the discovery process.

The second project is a systematic analysis of the regulatory framework of stand-alone MCPs. In organisms with large numbers of MCPs, it is probable that specific receptors are only expressed under certain environmental conditions and that the constituents of the chemoreceptor array change with time and environment. Associating MCPs with regulons by finding common regulatory motifs upstream of MCP sequences and other genes is a worthwhile project that would embed our understanding of chemotaxis function and evolution in an ecological, whole organism context, as well as provide interesting information for experimentalists interested in studying signal integration in the chemoreceptor array at the molecular level.

CHAPTER 7

CONCLUSION

If the genomics revolution dominated the last decade, the systems biology revolution will dominate the next decade or more. The research presented in this thesis rests on data from 312 complete and draft genome sequences, 152 of which have chemotaxis systems. The number of complete genome sequences now available is over 550 [129]. My hope is that genomic data will eventually saturate the phylogenetic tree, enable us to pinpoint each step change in the function of chemotaxis, and then step by step to trace its entire evolutionary history.

A brief review of some of the evolutionary steps studied in this thesis follows:

1) Diversity of lengths and methylation patterns in the MCP_CD domain imply that the chemotaxis signaling and adaptation mechanisms have co-evolved.

2) Loss of a key site of methylation in class 44H MCPs may have altered or destroyed the adaptation mechanism in 40H MCPs, leading to rewiring later in class 36H.

3) Loss of adaptation via the CheB / CheR pathway in *H. pylori* was compensated by an as yet undetermined mechanism that probably involved a reservoir of robustness encoded in CheV feedback.

4) In order to overcome their limited single-input architecture, two subfamilies of the Tfp / YWMA chemotaxis module followed divergent paths to increasing the complexity of their signaling mechanism. Cyanobacteria responded to the need for multiple inputs by simply duplicating the entire module, while proteobacteria responded by duplicating the number of phosphotransfer domains on the kinase.

The functional and evolutionary insights outlined above were possible because of (1) my development of computational tools to categorize a protein domain with a problematic structure and unique evolutionary mechanism, (2) my reliance on sequence analysis, aided by structure analysis, to extract important features from the resulting data, (3) my integration of information about receptor diversity with information about kinase diversity, again by development of a novel computational method, and (4) management and display of the data in an intuitive format that aids understanding and exploration of functional and evolutionary hypotheses.

In isolation, each of the examples outlined above is a set piece in evolution, explaining at the molecular level how a module transitioned from one functional regime to another in response to specific evolutionary pressures or to exploit particular selective advantages. The power of comparative genomics is not to see each example in isolation, but to integrate them into a unifying framework. Any elements of chemotaxis function that are universal can be better understood by applying constraints from data obtained from multiple organisms. A central conclusion of this research project is that it is no longer sufficient to focus on the mechanism of chemotaxis in *E. coli* and expect ever to gain a full understanding of the system. In the coming era of integrative systems biology, chemotaxis can serve as a model system for how to leverage legacy molecular biology experiments with comparative genomic data to guide future research towards problems of central interest.

APPENDIX A

Table A.1 GenBank accession numbers of all components of the 312 genomes examined in this study, including the date when the data was stored in GenBank. The FASTA abbreviation field is the species identifier that precedes each sequence identifier in multiple sequence alignments. Genomes are listed in alphabetical order with phylogenetic group indicated in parentheses. Abbreviations: WGS, Whole Genome Shotgun.
(alexander_roger_p_200712_phd_tableA1_accessions.pdf, 288 KB)

Figure A.1 Sequence logos of the seven major length classes of the MCP cytoplasmic domain. Heptad number and register are indicated at top. Alignment position number is indicated below the Class 44H sequence logo. Residues are colored as outlined in section 2.6.
(alexander_roger_p_200712_phd_figureA1_major_logo.pdf, 1.1 MB)

Figure A.2 Multiple sequence alignment of the seven major length classes of the MCP cytoplasmic domain in Stockholm format. Gaps between classes are indicated by dashes; gaps within classes by dots.
(alexander_roger_p_200712_phd_figureA2_major_aln.txt, 616 KB)

Figure A.3 Sequence logo of the alignment between parent class 38H and its children, 38+4H and 38+20H. Heptad number and register are indicated at top. Residues are colored as outlined in section 2.6.
(alexander_roger_p_200712_phd_figureA3_minor38H_logo.pdf, 2.1 MB)

Figure A.4 Multiple sequence alignment between parent class 38H and its children, 38+4H and 38+20H, in Stockholm format. Gaps between classes are indicated by dashes; gaps within classes by dots. The insertions in Classes 38+4H and 38+20H appear related, so aligning them to each other is meaningful.
(alexander_roger_p_200712_phd_figureA4_minor38H_aln.txt, 92 KB)

Figure A.5 Sequence logo of the alignment between parent class 40H and its children, 40+12H and 40+24H. Heptad number and register are indicated at top. Residues are colored as outlined in section 2.6.
(alexander_roger_p_200712_phd_figureA5_minor40H_logo.pdf, 2.1 MB)

Figure A.6 Multiple sequence alignment between parent class 40H and its child class 40+12H, in Stockholm format. Gaps between classes are indicated by dashes; gaps within classes by dots. The insertions in Classes 40+12H and 40+24H appear unrelated, so each must be aligned separately to the parent class 40H.
(alexander_roger_p_200712_phd_figureA6_minor40_12H_aln.txt, 276 KB)

Figure A.7 Multiple sequence alignment between parent class 40H and its child class 40+24H, in Stockholm format. Gaps between classes are indicated by dashes; gaps within classes by dots. The insertions in Classes 40+12H and 40+24H appear unrelated, so each must be aligned separately to the parent class 40H.
(alexander_roger_p_200712_phd_figureA7_minor40_24H_aln.txt, 332 KB)

Figure A.8 Sequence logo of the alignment between minor class 48H and its two possible parental classes, 44H and 36H. Heptad number and register are indicated at top. Residues are colored as outlined in section 2.6.
(alexander_roger_p_200712_phd_figureA8_minor48H_logo.pdf, 2.1 MB)

Figure A.9 Multiple sequence alignment between minor class 48H and its two possible parental classes, 44H and 36H, in Stockholm format. Gaps between classes are indicated by dashes; gaps within classes by dots.
(alexander_roger_p_200712_phd_figureA9_minor48H_aln.txt, 328 KB)

REFERENCES

1. Avery, O.T., Macleod, C.M. & McCarty, M. Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types - Induction of Transformation by a Deoxyribonucleic-Acid Fraction Isolated From Pneumococcus Type-III. *J Exp Med* **79**, 137-158 (1944).

2. Hershey, A.D. & Chase, M. Independent functions of viral protein and nucleic acid in growth of bacteriophage. *J Gen Physiol* **36**, 39-56 (1952).

3. Watson, J.D. & Crick, F.H.C. Molecular Structure of Nucleic Acids - A Structure for Deoxyribose Nucleic Acid. *Nature* **171**, 737-738 (1953).

4. Alon, U. Simplicity in biology. *Nature* **446**, 497-497 (2007).

5. Hartwell, L.H., Hopfield, J.J., Leibler, S. & Murray, A.W. From molecular to modular cell biology. *Nature* **402**, C47-C52 (1999).

6. Kashtan, N. & Alon, U. Spontaneous evolution of modularity and network motifs. *Proc Natl Acad Sci USA* **102**, 13773-13778 (2005).

7. Adler, J. Chemotaxis in bacteria. *Science* **153**, 708-16 (1966).

8. Armitage, J.P. Bacterial tactic responses. *Adv Microb Physiol* **41**, 229-289 (1999).

9. Berg, H.C. & Brown, D.A. Chemotaxis in Escherichia-Coli Analyzed by 3-Dimensional Tracking. *Nature* **239**, 500-& (1972).

10. Bardy, S.L., Ng, S.Y.M. & Jarrell, K.F. Prokaryotic motility structures. *Microbiol (UK)* **149**, 295-304 (2003).

11. Schmitt, R. Sinorhizobial chemotaxis: a departure from the enterobacterial paradigm. *Microbiol (UK)* **148**, 627-631 (2002).

12. Armitage, J.P. & Schmitt, R. Bacterial chemotaxis: Rhodobacter sphaeroides and Sinorhizobium meliloti - variations on a theme? *Microbiol (UK)* **143**, 3671-3682 (1997).

13. Taylor, B.L. & Koshland, D.E. Reversal of Flagellar Rotation in Monotrichous and Peritrichous Bacteria - Generation of Changes in Direction. *J Bacteriol* **119**, 640-642 (1974).

14. Charon, N.W. & Goldstein, S.F. Genetics of motility and chemotaxis of a fascinating group of bacteria: The spirochetes. *Annu Rev Genet* **36**, 47-73 (2002).

15. Skerker, J.M. & Berg, H.C. Direct observation of extension and retraction of type IV pili. *Proc Natl Acad Sci USA* **98**, 6901-6904 (2001).

16. Harshey, R.M. Bees Arent the Only Ones - Swarming in Gram-Negative Bacteria. *Mol Microbiol* **13**, 389-394 (1994).

17. Harshey, R.M. & Matsuyama, T. Dimorphic Transition in Escherichia-Coli and Salmonella-Typhimurium - Surface-Induced Differentiation into Hyperflagellate Swarmer Cells. *Proc Natl Acad Sci USA* **91**, 8631-8635 (1994).

18. Burkart, M., Toguchi, A. & Harshey, R.M. The chemotaxis system, but not chemotaxis, is essential for swarming motility in Escherichia coli. *Proc Natl Acad Sci USA* **95**, 2568-2573 (1998).

19. Berleman, J.E. & Bauer, C.E. A che-like signal transduction cascade involved in controlling flagella biosynthesis in Rhodospirillum centenum. *Mol Microbiol* **55**, 1390-1402 (2005).

20. Berleman, J.E. & Bauer, C.E. Involvement of a Che-like signal transduction cascade in regulating cyst cell development in Rhodospirillum centenum. *Mol Microbiol* **56**, 1457-1466 (2005).

21. Kirby, J.R. & Zusman, D.R. Chemosensory regulation of developmental gene expression in Myxococcus xanthus. *Proc Natl Acad Sci USA* **100**, 2008-2013 (2003).

22. Hickman, J.W., Tifrea, D.F. & Harwood, C.S. A chemosensory system that regulates biofilm formation through modulation of cyclic diguanylate levels. *Proc Natl Acad Sci USA* **102**, 14422-14427 (2005).

23. Koretke, K.K., Lupas, A.N., Warren, P.V., Rosenberg, M. & Brown, J.R. Evolution of two-component signal transduction. *Mol Biol Evol* **17**, 1956-1970 (2000).

24. Ulrich, L.E., Koonin, E.V. & Zhulin, I.B. One-component systems dominate signal transduction in prokaryotes. *Trends Microbiol* **13**, 52-56 (2005).

25. Wuichet, K. Ph.D. Thesis, Georgia Institute of Technology (2007).

26. Stock, A.M., Robinson, V.L. & Goudreau, P.N. Two-component signal transduction. *Annu Rev Biochem* **69**, 183-215 (2000).

27. Dutta, R., Qin, L. & Inouye, M. Histidine kinases: diversity of domain organization. *Mol Microbiol* **34**, 633-640 (1999).

28. Adler, J. & Tso, W.W. Decision-Making in Bacteria - Chemotactic Response of Escherichia-Coli to Conflicting Stimuli. *Science* **184**, 1293-1294 (1974).

29. Tso, W.W. & Adler, J. Negative Chemotaxis in Escherichia coli. *J Bacteriol* **118**, 560-576 (1974).

30. Khan, S., Spudich, J.L., Mccray, J.A. & Trentham, D.R. Chemotactic Signal Integration in Bacteria. *Proc Natl Acad Sci USA* **92**, 9757-9761 (1995).

31. Maddock, J.R. & Shapiro, L. Polar Location of the Chemoreceptor Complex in the Escherichia-Coli Cell. *Science* **259**, 1717-1723 (1993).

32. Gestwicki, J.E., Lamanna, A.C., Harshey, R.M., McCarter, L.L., Kiessling, L.L. & Adler, J. Evolutionary conservation of methyl-accepting chemotaxis protein location in Bacteria and Archaea. *J Bacteriol* **182**, 6499-6502 (2000).

33. Sourjik, V. & Berg, H.C. Functional interactions between receptors in bacterial chemotaxis. *Nature* **428**, 437-41 (2004).

34. Ames, P., Studdert, C.A., Reiser, R.H. & Parkinson, J.S. Collaborative signaling by mixed chemoreceptor teams in Escherichia coli. *Proc Natl Acad Sci USA* **99**, 7060-7065 (2002).

35. Studdert, C.A. & Parkinson, J.S. Crosslinking snapshots of bacterial chemoreceptor squads. *Proc Natl Acad Sci USA* **101**, 2117-2122 (2004).

36. Zhang, P., Khursigara, C.M., Hartnell, L.M. & Subramaniam, S. Direct visualization of Escherichia coli chemotaxis receptor arrays using cryo-electron microscopy. *Proc Natl Acad Sci USA* **104**, 3777-3781 (2007).

37. Alexander, R.P. & Zhulin, I.B. Evolutionary genomics reveals conserved structural determinants of signaling and adaptation in microbial chemoreceptors. *Proc Natl Acad Sci USA* **104**, 2885-2890 (2007).

38. Segall, J.E., Block, S.M. & Berg, H.C. Temporal comparisons in bacterial chemotaxis. *Proc Natl Acad Sci USA* **83**, 8987-91 (1986).

39. Wang, H. & Matsumura, P. Characterization of the CheA(S)/CheZ complex: A specific interaction resulting in enhanced dephosphorylating activity on CheY-phosphate. *Mol Microbiol* **19**, 695-703 (1996).

40. Zhao, R., Collins, E.J., Bourret, R.B. & Silversmith, R.E. Structure and catalytic mechanism of the E. coli chemotaxis phosphatase CheZ. **9**, 570-5 (2002).

41. Lipkow, K., Andrews, S.S. & Bray, D. Simulated diffusion of phosphorylated CheY through the cytoplasm of Escherichia coli. *J Bacteriol* **187**, 45-53 (2005).

42. Rao, C.V., Kirby, J.R. & Arkin, A.P. Phosphatase localization in bacterial chemotaxis: divergent mechanisms, convergent principles. *Phys Biol* **2**, 148 (2005).

43. Schubert, H.L., Blumenthal, R.M. & Cheng, X.D. Many paths to methyltransfer: a chronicle of convergence. *Trends Biochem Sci* **28**, 329-335 (2003).

44. Shiomi, D., Zhulin, I.B., Homma, M. & Kawagishi, I. Dual recognition of the bacterial chemoreceptor by chemotaxis-specific domains of the CheR methyltransferase. *J Biol Chem* **277**, 42325-42333 (2002).

45. Djordjevic, S. & Stock, A.M. Chemotaxis receptor recognition by protein methyltransferase CheR. *Nat Struct Biol* **5**, 446-450 (1998).

46. Perez, E., West, A.H., Stock, A.M. & Djordjevic, S. Discrimination between different methylation states of chemotaxis receptor tar by receptor methyltransferase CheR. *Biochemistry* **43**, 953-961 (2004).

47. West, A.H., Martinezhackert, E. & Stock, A.M. Crystal-Structure of the Catalytic Domain of the Chemotaxis Receptor Methylesterase, Cheb. *J Mol Biol* **250**, 276-290 (1995).

48. Anand, G.S., Goudreau, P.N. & Stock, A.M. Activation of methylesterase CheB: Evidence of a dual role for the regulatory domain. *Biochemistry* **37**, 14038-14047 (1998).

49. Djordjevic, S., Goudreau, P.N., Xu, Q.P., Stock, A.M. & West, A.H. Structural basis for methylesterase CheB regulation by a phosphorylation-activated domain. *Proc Natl Acad Sci USA* **95**, 1381-1386 (1998).

50. Falke, J.J. & Hazelbauer, G.L. Transmembrane signaling in bacterial chemoreceptors. *Trends Biochem Sci* **26**, 257-265 (2001).

51. Kehry, M.R., Bond, M.W., Hunkapiller, M.W. & Dahlquist, F.W. Enzymatic Deamidation of Methyl-Accepting Chemotaxis Proteins in Escherichia-Coli Catalyzed by the CheB Gene-Product. *Proc Natl Acad Sci USA* **80**, 3599-3603 (1983).

52. Goldbeter, A. & Koshland, D.E. An Amplified Sensitivity Arising from Covalent Modification in Biological-Systems. *Proc Natl Acad Sci USA* **78**, 6840-6844 (1981).

53. Anand, G.S. & Stock, A.M. Kinetic basis for the stimulatory effect of phosphorylation on the methylesterase activity of CheB. *Biochemistry* **41**, 6752-6760 (2002).

54. Berg, H.C. The rotary motor of bacterial flagella. *Annu Rev Biochem* **72**, 19-54 (2003).

55. Rosario, M.M.L., Kirby, J.R., Bochar, D.A. & Ordal, G.W. Chemotactic Methylation and Behavior in Bacillus Subtilis - Role of 2 Unique Proteins, CheC and CheD. *Biochemistry* **34**, 3823-3831 (1995).

56. Alon, U., Camarena, L., Surette, M.G., Arcas, B.A.Y., Liu, Y., Leibler, S. & Stock, J.B. Response regulator output in bacterial chemotaxis. *Embo J* **17**, 4238-4248 (1998).

57. Sourjik, V. & Berg, H.C. Receptor sensitivity in bacterial chemotaxis. *Proc Natl Acad Sci USA* **99**, 123-127 (2002).

58. Barkai, N. & Leibler, S. Robustness in simple biochemical networks. *Nature* **387**, 913-917 (1997).

59. Alon, U., Surette, M.G., Barkai, N. & Leibler, S. Robustness in bacterial chemotaxis. *Nature* **397**, 168-171 (1999).

60. Morton-Firth, C.J., Shimizu, T.S. & Bray, D. A free-energy-based stochastic simulation of the Tar receptor complex. *J Mol Biol* **286**, 1059-1074 (1999).

61. Yi, T.M., Huang, Y., Simon, M.I. & Doyle, J. Robust perfect adaptation in bacterial chemotaxis through integral feedback control. *Proc Natl Acad Sci USA* **97**, 4649-4653 (2000).

62. Mello, B.A. & Tu, Y. Perfect and near-perfect adaptation in a model of bacterial chemotaxis. *Biophys J* **84**, 2943-2956 (2003).

63. Shimizu, T.S., Aksenov, S.V. & Bray, D. A spatially extended stochastic model of the bacterial chemotaxis signalling pathway. *J Mol Biol* **329**, 291-309 (2003).

64. Korobkova, E., Emonet, T., Vilar, J.M.G., Shimizu, T.S. & Cluzel, P. From molecular noise to behavioural variability in a single bacterium. *Nature* **428**, 574-578 (2004).

65. Rao, C.V., Kirby, J.R. & Arkin, A.P. Design and Diversity in Bacterial Chemotaxis: A Comparative Study in Escherichia coli and Bacillus subtilis. *PLoS Biology* **2**, 239-252 (2004).

66. Keymer, J.E., Endres, R.G., Skoge, M., Meir, Y. & Wingreen, N.S. Chemosensing in Escherichia coli: Two regimes of two-state receptors. *Proc Natl Acad Sci USA* **103**, 1786-1791 (2006).

67. Kollmann, M., Lovdok, L., Bartholome, K., Timmer, J. & Sourjik, V. Design principles of a bacterial signalling network. *Nature* **438**, 504-507 (2005).

68. Chu, S. Invited Lecture, Georgia Tech School of Biomedical Engineering. (2006).

69. Thomas, D.R., Francis, N.R., Xu, C. & DeRosier, D.J. The three-dimensional structure of the flagellar rotor from a clockwise-locked mutant of Salmonella enterica serovar typhimurium. *J Bacteriol* **188**, 7039-7048 (2006).

70. Cluzel, P., Surette, M. & Leibler, S. An ultrasensitive bacterial motor revealed by monitoring signaling proteins in single cells. *Science* **287**, 1652-1655 (2000).

71. Kirby, J.R., Kristich, C.J., Saulmon, M.M., Zimmer, M.A., Garrity, L.F., Zhulin, I.B. & Ordal, G.W. CheC is related to the family of flagellar switch proteins and acts independently from CheD to control chemotaxis in Bacillus subtilis. *Mol Microbiol* **42**, 573-585 (2001).

72. Szurmant, H., Muff, T.J. & Ordal, G.W. Bacillus subtilis CheC and FliY are members of a novel class of CheY-P-hydrolyzing proteins in the chemotactic signal transduction cascade. *J Biol Chem* **279**, 21787-21792 (2004).

73. Saulmon, M.M., Karatan, E. & Ordal, G.W. Effect of loss of CheC and other adaptational proteins on chemotactic behaviour in Bacillus subtilis. *Microbiol (UK)* **150**, 581-589 (2004).

74. Park, S.Y., Chao, X.J., Gonzalez-Bonet, G., Beel, B.D., Bilwes, A.M. & Crane, B.R. Structure and function of an unusual family of protein phosphatases: The bacterial chemotaxis proteins CheC and CheX. *Mol Cell* **16**, 563-574 (2004).

75. Karatan, E., Saulmon, M.M., Bunn, M.W. & Ordal, G.W. Phosphorylation of the response regulator CheV is required for adaptation to attractants during Bacillus subtilis chemotaxis. *J Biol Chem* **276**, 43618-43626 (2001).

76. Kristich, C.J. & Ordal, G.W. Bacillus subtilis CheD is a chemoreceptor modification enzyme required for chemotaxis. *J Biol Chem* **277**, 25356-25362 (2002).

77. Park, C.Y., Dutton, D.P. & Hazelbauer, G.L. Effects of Glutamines and Glutamates at Sites of Covalent Modification of a Methyl-Accepting Transducer. *J Bacteriol* **172**, 7179-7187 (1990).

78. Rice, M.S. & Dahlquist, F.W. Sites of Deamidation and Methylation in Tsr, a Bacterial Chemotaxis Sensory Transducer. *J Biol Chem* **266**, 9746-9753 (1991).

79. Barnakov, A.N., Barnakova, L.A. & Hazelbauer, G.L. Efficient adaptational demethylation of chemoreceptors requires the same enzyme-docking site as efficient methylation. *Proc Natl Acad Sci USA* **96**, 10667-10672 (1999).

80. Rosario, M.M.L. & Ordal, G.W. CheC and CheD interact to regulate methylation of Bacillus subtilis methyl-accepting chemotaxis proteins. *Mol Microbiol* **21**, 511-518 (1996).

81. Chao, X., Muff, T.J., Park, S.-Y., Zhang, S., Pollard, A.M., Ordal, G.W., Bilwes, A.M. & Crane, B.R. A Receptor-Modifying Deamidase in Complex with a Signaling Phosphatase Reveals Reciprocal Regulation. *Cell* **124**, 561-571 (2006).

82. Soyer, O.S. & Bonhoeffer, S. Evolution of complexity in signaling pathways. *Proc Natl Acad Sci U S A* **103**, 16337-16342 (2006).

83. Zhulin, I.B. The superfamily of chemotaxis transducers: from physiology to genomics and back. *Adv Micro Physiol* **45**, 157-98 (2001).

84. Ulrich, L.E. & Zhulin, I.B. Four-helix bundle: a ubiquitous sensory module in prokaryotic signal transduction. *Bioinformatics* **21**, iii45-48 (2005).

85. Ma, Q.H., Johnson, M.S. & Taylor, B.L. Genetic analysis of the HAMP domain of the Aer aerotaxis sensor localizes flavin adenine dinucleotide-binding determinants to the AS-2 helix. *J Bacteriol* **187**, 193-201 (2005).

86. Brooun, A., Bell, J., Freitas, T., Larsen, R.W. & Alam, M. An archaeal aerotaxis transducer combines subunit I core structures of eukaryotic cytochrome c oxidase and eubacterial methyl-accepting chemotaxis proteins. *J Bacteriol* **180**, 1642-1646 (1998).

87. Wuichet, K. & Zhulin, I.B. Molecular evolution of sensory domains in cyanobacterial chemoreceptors. *Trends Microbiol* **11**, 200-203 (2003).

88. Zhang, W. & Phillips, G.N. Structure of the oxygen sensor in Bacillus subtilis: Signal transduction of chemotaxis by control of symmetry. *Structure* **11**, 1097-1110 (2003).

89. LeMoual, H. & Koshland, D.E. Molecular evolution of the C-terminal cytoplasmic domain of a superfamily of bacterial receptors involved in taxis. *J Mol Biol* **261**, 568-585 (1996).

90. Kim, K.K., Yokota, H. & Kim, S.H. Four-helical-bundle structure of the cytoplasmic domain of a serine chemotaxis receptor. *Nature* **400**, 787-792 (1999).

91. Crick, F.H.C. The Packing of Alpha-Helices - Simple Coiled-Coils. *Acta Crystallogr* **6**, 689-697 (1953).

92. Walshaw, J. & Woolfson, D.N. Socket: a program for identifying and analysing coiled-coil motifs within protein structures. *J Mol Biol* **307**, 1427-1450 (2001).

93. Walshaw, J. & Woolfson, D.N. Extended knobs-into-holes packing in classical and complex coiled-coil assemblies. *J Struct Biol* **144**, 349-361 (2003).

94. Park, S.Y., Borbat, P.P., Gonzalez-Bonet, G., Bhatnagar, J., Pollard, A.M., Freed, J.H., Bilwes, A.M. & Crane, B.R. Reconstruction of the chemotaxis receptor-kinase assembly. *Nat Struct Molec Biol* **13**, 400-407 (2006).

95. Aravind, L. & Ponting, C.P. The cytoplasmic helical linker domain of receptor histidine kinase and methyl-accepting proteins is common to many prokaryotic signalling proteins. *FEMS Microbiol Lett* **176**, 111-116 (1999).

96. Danielson, M.A., Bass, R.B. & Falke, J.J. Cysteine and disulfide scanning reveals a regulatory alpha-helix in the cytoplasmic domain of the aspartate receptor. *J Biol Chem* **272**, 32878-32888 (1997).

97. Butler, S.L. & Falke, J.J. Cysteine and disulfide scanning reveals two amphiphilic helices in the linker region of the aspartate chemoreceptor. *Biochemistry* **37**, 10746-10756 (1998).

98. Singh, M., Berger, B., Kim, P.S., Berger, J.M. & Cochran, A.G. Computational learning reveals coiled coil-like motifs in histidine kinase linker domains. *Proc Natl Acad Sci USA* **95**, 2738-2743 (1998).

99. Williams, S.B. & Stewart, V. Functional similarities among two-component sensors and methyl-accepting chemotaxis proteins suggest a role for linker region amphipathic helices in transmembrane signal transduction. *Mol Microbiol* **33**, 1093-1102 (1999).

100. Chen, X.P. & Spudich, J.L. Five residues in the HtrI transducer membrane-proximal domain close the cytoplasmic proton-conducting channel of sensory rhodopsin. *J Biol Chem* **279**, 42964-42969 (2004).

101. Kristich, C.J. & Ordal, G.W. Analysis of chimeric chemoreceptors in Bacillus subtilis reveals a role for CheD in the function of the McpC HAMP domain. *J Bacteriol* **186**, 5950-5955 (2004).

102. Watts, K.J., Ma, Q.H., Johnson, M.S. & Taylor, B.L. Interactions between the PAS and HAMP domains of the Escherichia coli aerotaxis receptor Aer. *J Bacteriol* **186**, 7440-7449 (2004).

103. Yang, C.S., Sineshchekov, O., Spudich, E.N. & Spudich, J.L. The cytoplasmic membrane-proximal domain of the HtrII transducer interacts with the E-F loop of photoactivated Natronomonas pharaonis sensory rhodopsin II. *J Biol Chem* **279**, 42970-42976 (2004).

104. Bordignon, E., Klare, J.P., Doebber, M., Wegener, A.A., Martell, S., Engelhard, M. & Steinhoff, H.J. Structural analysis of a HAMP domain - The linker region of the phototransducer in complex with sensory rhodopsin II. *J Biol Chem* **280**, 38767-38775 (2005).

105. Sudo, Y., Okuda, H., Yamabi, M., Fukuzaki, Y., Mishima, M., Kamo, N. & Kojima, C. Linker region of a halobacterial transducer protein interacts directly with its sensor retinal protein. *Biochemistry* **44**, 6144-6152 (2005).

106. Buron-Barral, M.D., Gosink, K.K. & Parkinson, J.S. Loss- and gain-of-function mutations in the F1-HAMP region of the Escherichia coli aerotaxis transducer Aer. *J Bacteriol* **188**, 3477-3486 (2006).

107. Hulko, M. et al. The HAMP domain structure implies helix rotation in transmembrane signaling. *Cell* **126**, 929-940 (2006).

108. Taylor, B.L. & Zhulin, I.B. PAS domains: Internal sensors of oxygen, redox potential, and light. *Microbiol Mol Biol Rev* **63**, 479-506 (1999).

109. Aravind, L. & Ponting, C.P. The GAF domain: an evolutionary link between diverse phototransducing proteins. *Trends BiochemSci* **22**, 458-459 (1997).

110. Reinelt, S., Hofmann, E., Gerharz, T., Bott, M. & Madden, D.R. The structure of the periplasmic ligand-binding domain of the sensor kinase CitA reveals the first extracellular PAS domain. *J Biol Chem* **278**, 39189-39196 (2003).

111. Anantharaman, V. & Aravind, L. Cache - a signaling domain common to animal Ca2+ channel subunits and a class of prokaryotic chemotaxis receptors. *Trends BiochemSci* **25**, 535-537 (2000).

112. Shu, C.J., Ulrich, L.E. & Zhulin, I.B. The NIT domain: a predicted nitrate-responsive module in bacterial sensory receptors. *Trends BiochemSci* **28**, 121-124 (2003).

113. Chervitz, S.A. & Falke, J.J. Molecular mechanism of transmembrane signaling by the aspartate receptor: A model. *Proc Natl Acad Sci USA* **93**, 2545-2550 (1996).

114. Draheim, R.R., Bormans, A.F., Lai, R.Z. & Manson, M.D. Tuning a bacterial chemoreceptor with protein-membrane interactions. *Biochemistry* **45**, 14655-14664 (2006).

115. Bass, R.B. & Falke, J.J. The aspartate receptor cytoplasmic domain: in situ chemical analysis of structure, mechanism and dynamics. *Struct Fold Des* **7**, 829-840 (1999).

116. Klenk, H.P. et al. The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon Archaeoglobus fulgidus. *Nature* **390**, 364-370 (1997).

117. Moukhametzianov, R., Klare, J.P., Efremov, R., Baeken, C., Goppner, A., Labahn, J., Engelhard, M., Buldt, G. & Gordeliy, V.I. Development of the signal in sensory rhodopsin and its transfer to the cognate transducer. *Nature* **440**, 115-119 (2006).

118. Szurmant, L. & Ordal, G.W. Diversity in chemotaxis mechanisms among the bacteria and archaea. *Microbiol Mol Biol Rev* **68**, 301-+ (2004).

119. Bischoff, D.S., Bourret, R.B., Kirsch, M.L. & Ordal, G.W. Purification and Characterization of Bacillus-Subtilis CheY. *Biochemistry* **32**, 9256-9261 (1993).

120. Szurmant, H., Bunn, M.W., Cannistraro, V.J. & Ordal, G.W. Bacillus subtilis hydrolyzes CheY-P at the location of its action, the flagellar switch. *J Biol Chem* **278**, 48611-48616 (2003).

121. Zimmer, M.A., Tiu, J., Collins, M.A. & Ordal, G.T. Selective methylation changes on the Bacillus subtilis chemotaxis receptor McpB promote adaptation. *J Biol Chem* **275**, 24264-24272 (2000).

122. Wheeler, D.L. et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **35**, D5-D12 (2007).

123. Pruitt, K.D., Tatusova, T. & Maglott, D.R. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **35**, D61-D65 (2007).

124. Boeckmann, B. et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* **31**, 365-370 (2003).

125. Ulrich, L.E. & Zhulin, I.B. MiST: a microbial signal transduction database. *Nucleic Acids Res* **35**, D386-D390 (2007).

126. Ulrich, L.E. Ph.D. Thesis, Georgia Institute of Technology (2005).

127. Finn, R.D. et al. Pfam: clans, web tools and services. *Nucleic Acids Res* **34**, D247-251 (2006).

128. Letunic, I., Copley, R.R., Pils, B., Pinkert, S., Schultz, J. & Bork, P. SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res* **34**, D257-D260 (2006).

129. Completed Microbial Genomes. Retrieved 14 August, 2007, from ftp://ftp.ncbi.nih.gov/genomes/Bacteria/Complete.txt.

130. Sonnhammer, E.L.L., Eddy, S.R. & Durbin, R. Pfam: A comprehensive database of protein domain families based on seed alignments. *Proteins* **28**, 405-420 (1997).

131. Schultz, J., Milpetz, F., Bork, P. & Ponting, C.P. SMART, a simple modular architecture research tool: Identification of signaling domains. *Proc Natl Acad Sci USA* **95**, 5857-5864 (1998).

132. Tatusov, R.L., Koonin, E.V. & Lipman, D.J. A genomic perspective on protein families. *Science* **278**, 631-637 (1997).

133. Tatusov, R.L., Galperin, M.Y., Natale, D.A. & Koonin, E.V. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* **28**, 33-36 (2000).

134. Berman, H., Henrick, K. & Nakamura, H. Announcing the worldwide Protein Data Bank. *Nat Struct Biol* **10**, 980-980 (2003).

135. Dayhoff, M.O., Schwartz, R.M. & Orcutt, B.C. A model of evolutionary change in proteins. *Atlas of Protein Sequence and Structure* **5**, 345-352 (1978).

136. Henikoff, S. & Henikoff, J.G. Amino-Acid Substitution Matrices from Protein Blocks. *Proc Natl Acad Sci USA* **89**, 10915-10919 (1992).

137. Needleman, S.B. & Wunsch, C.D. A General Method Applicable to Search for Similarities in Amino Acid Sequence of Two Proteins. *J Mol Biol* **48**, 443-453 (1970).

138. Smith, T.F. & Waterman, M.S. Identification of Common Molecular Subsequences. *J Mol Biol* **147**, 195-197 (1981).

139. Feng, D.F. & Doolittle, R.F. Progressive Sequence Alignment as a Prerequisite to Correct Phylogenetic Trees. *J Mol Evol* **25**, 351-360 (1987).

140. Durbin, R., Eddy, S., Krogh, A. & Mitchison, G. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, Cambridge (1998).

141. Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T.J., Higgins, D.G. & Thompson, J.D. Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res* **31**, 3497-3500 (2003).

142. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic Local Alignment Search Tool. *J Mol Biol* **215**, 403-410 (1990).

143. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J.H., Zhang, Z., Miller, W. & Lipman, D.J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-3402 (1997).

144. Altschul, S.F. & Koonin, E.V. Iterated profile searches with PSI-BLAST - a tool for discovery in protein databases. *Trends BiochemSci* **23**, 444-447 (1998).

145. Schaffer, A.A., Aravind, L., Madden, T.L., Shavirin, S., Spouge, J.L., Wolf, Y.I., Koonin, E.V. & Altschul, S.F. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res* **29**, 2994-3005 (2001).

146. Eddy, S.R. Hidden Markov models. *Curr Opin Struc Biol* **6**, 361-365 (1996).

147. Eddy, S.R. Profile hidden Markov models. *Bioinformatics* **14**, 755-763 (1998).

148. Eddy, S.R. What is a hidden Markov model? *Nat Biotechnol* **22**, 1315-1316 (2004).

149. Coin, L., Bateman, A. & Durbin, R. Enhanced protein domain discovery by using language modeling techniques from speech recognition. *Proc Natl Acad Sci USA* **100**, 4516-4520 (2003).

150. Price, M.N., Arkin, A.P. & Alm, E.J. The life-cycle of operons. *PLoS Genet* **2**, 859-873 (2006).

151. Price, M.N., Huang, K.H., Alm, E.J. & Arkin, A.P. A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic Acids Res* **33**, 880-892 (2005).

152. Shannon, C.E. A Mathematical Theory of Communication. *Bell Systems Tech J* **27**, 379-423, 623-656 (1948).

153. Schneider, T.D. & Stephens, R.M. Sequence Logos - a New Way to Display Consensus Sequences. *Nucleic Acids Res* **18**, 6097-6100 (1990).

154. Crooks, G.E., Hon, G., Chandonia, J.M. & Brenner, S.E. WebLogo: A sequence logo generator. *Genome Res* **14**, 1188-1190 (2004).

155. Kumar, S., Tamura, K. & Nei, M. MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment. *Brief Bioinform* **5**, 150-163 (2004).

156. Felsenstein, J. *Inferring Phylogenies*, Sinauer Associates, Sunderland, Massachusetts (2004).

157. Saitou, N. & Nei, M. The Neighbor-Joining Method - a New Method for Reconstructing Phylogenetic Trees. *Mol Biol Evol* **4**, 406-425 (1987).

158. Tamura, K., Nei, M. & Kumar, S. Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proc Natl Acad Sci USA* **101**, 11030-11035 (2004).

159. Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52**, 696-704 (2003).

160. Kail, L., Krogh, A. & Sonnhammer, E.L.L. A combined transmembrane topology and signal peptide prediction method. *J Mol Biol* **338**, 1027-1036 (2004).

161. Fiser, A.S. & Sali, A. MODELLER: Generation and refinement of homology-based protein structure models. *Methods Enzymol* **374**, 461-491 (2003).

162. Cserzo, M., Eisenhaber, F., Eisenhaber, B. & Simon, I. On filtering false positive transmembrane protein predictions. *Protein Eng* **15**, 745-752 (2002).

163. Terwilliger, T.C., Wang, J.Y. & Koshland, D.E. Kinetics of Receptor Modification - the Multiply Methylated Aspartate Receptors Involved in Bacterial Chemotaxis. *J Biol Chem* **261**, 814-820 (1986).

164. Koch, M.K. & Oesterhelt, D. MpcT is the transducer for membrane potential changes in Halobacterium salinarum. *Mol Microbiol* **55**, 1681-1694 (2005).

165. Perazzona, B. & Spudich, J.L. Identification of methylation sites and effects of phototaxis stimuli on transducer methylation in Halobacterium salinarum. *J Bacteriol* **181**, 5676-5683 (1999).

166. Perez, E., Zheng, H. & Stock, A.M. Identification of methylation sites in Thermotoga maritima chemotaxis receptors. *J Bacteriol* **188**, 4093-4100 (2006).

167. Allwood, A.C., Walter, M.R., Kamber, B.S., Marshall, C.P. & Burch, I.W. Stromatolite reef from the Early Archaean era of Australia. *Nature* **441**, 714-718 (2006).

168. Battistuzzi, F.U., Feijao, A. & Hedges, S.B. A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colonization of land. *BMC Evol Biol* **4**, 14 (2004).

169. Gernert, K.M., Surles, M.C., Labean, T.H., Richardson, J.S. & Richardson, D.C. The Alacoil - a Very Tight, Antiparallel Coiled-Coil of Helices. *Protein Sci* **4**, 2252-2260 (1995).

170. Kim, S.H., Wang, W. & Kim, K.K. Dynamic and clustering model of bacterial chemotaxis receptors: structural basis for signaling and high sensitivity. *Proc Natl Acad Sci USA* **99**, 11611-11615 (2002).

171. Coleman, M.D., Bass, R.B., Mehan, R.S. & Falke, J.J. Conserved glycine residues in the cytoplasmic domain of the aspartate receptor play essential roles in kinase coupling and on-off switching. *Biochemistry* **44**, 7687-7695 (2005).

172. Shapiro, M.J., Panomitros, D. & Koshland, D.E. Interactions between the Methylation Sites of the Escherichia-Coli Aspartate Receptor-Mediated by the Methyltransferase. *J Biol Chem* **270**, 751-755 (1995).

173. Alam, M., Lebert, M., Oesterhelt, D. & Hazelbauer, G.L. Methyl-Accepting Taxis Proteins in Halobacterium-Halobium. *Embo J* **8**, 631-639 (1989).

174. Zhang, W.S., Brooun, A., McCandless, J., Banda, P. & Alam, M. Signal transduction in the Archaeon Halobacterium salinarium is processed through three subfamilies of 13 soluble and membrane-bound transducer proteins. *Proc Natl Acad Sci USA* **93**, 4649-4654 (1996).

175. Brooun, A., Zhang, W.S. & Alam, M. Primary structure and functional analysis of the soluble transducer protein HtrXI in the archaeon Halobacterium salinarium. *J Bacteriol* **179**, 2963-2968 (1997).

176. Hou, S.B., Brooun, A., Yu, H.S., Freitas, T. & Alam, M. Sensory rhodopsin II transducer HtrII is also responsible for serine chemotaxis in the archaeon Halobacterium salinarum. *J Bacteriol* **180**, 1600-1602 (1998).

177. Storch, K.F., Rudolph, J. & Oesterhelt, D. Car: a cytoplasmic sensor responsible for arginine chemotaxis in the archaeon Halobacterium salinarum. *Embo J* **18**, 1146-1158 (1999).

178. Kokoeva, M.V. & Oesterhelt, D. BasT, a membrane-bound transducer protein for amino acid detection in Halobacterium salinarum. *Mol Microbiol* **35**, 647-656 (2000).

179. Kokoeva, M.V., Storch, K.F., Klein, C. & Oesterhelt, D. A novel mode of sensory transduction in archaea: binding protein-mediated chemotaxis towards osmoprotectants and amino acids. *Embo J* **21**, 2312-2322 (2002).

180. Ames, P. & Parkinson, J.S. Conformational suppression of inter-receptor signaling defects. *Proc Natl Acad Sci USA* **103**, 9292-9297 (2006).

181. Studdert, C.A. & Parkinson, J.S. Insights into the organization and dynamics of bacterial chemoreceptor clusters through in vivo crosslinking studies. *Proc Natl Acad Sci USA* **102**, 15623-15628 (2005).

182. Francis, N.R., Wolanin, P.M., Stock, J.B., DeRosier, D.J. & Thomas, D.R. Three-dimensional structure and organization of a receptor/signaling complex. *Proc Natl Acad Sci USA* **101**, 17480-17485 (2004).

183. Asakura, S. & Honda, H. Two-state model for bacterial chemoreceptor proteins: The role of multiple methylation. *J Mol Biol* **176**, 349-367 (1984).

184. Bornhorst, J.A. & Falke, J.J. Quantitative analysis of aspartate receptor signaling complex reveals that the homogeneous two-state model is inadequate: development of a heterogeneous two-state model. *J Mol Biol* **326**, 1597-1614 (2003).

185. Crane, B.R. Personal Communication. (2007).

186. Homma, M., Shiomi, D. & Kawagishi, I. Attractant binding alters arrangement of chemoreceptor dimers within its cluster at a cell pole. *Proc Natl Acad Sci USA* **101**, 3462-3467 (2004).

187. Bray, D., Levin, M.D. & Morton-Firth, C.J. Receptor clustering as a cellular mechanism to control sensitivity. *Nature* **393**, 85-88 (1998).

188. Duke, T.A.J. & Bray, D. Heightened sensitivity of a lattice of membrane receptors. *Proc Natl Acad Sci USA* **96**, 10104-10108 (1999).

189. Liberman, L., Berg, H.C. & Sourjik, V. Effect of chemoreceptor modification on assembly and activity of the receptor-kinase complex in Escherichia coli. *J Bacteriol* **186**, 6643-6646 (2004).

190. Vaknin, A. & Berg, H.C. Osmotic stress mechanically perturbs chemoreceptors in Escherichia coli. *Proc Natl Acad Sci USA* **103**, 592-596 (2006).

191. Li, M.S. & Hazelbauer, G.L. The carboxyl-terminal linker is important for chemoreceptor function. *Mol Microbiol* **60**, 469-479 (2006).

192. Li, M.S. & Hazelbauer, G.L. Adaptational assistance in clusters of bacterial chemoreceptors. *Mol Microbiol* **56**, 1617-1626 (2005).

193. Endres, R.G. & Wingreen, N.S. Precise adaptation in bacterial chemotaxis through "assistance neighborhoods". *Proc Natl Acad Sci USA* **103**, 13040-13044 (2006).

194. Guvener, Z.T., Tifrea, D.F. & Harwood, C.S. Two different Pseudomonas aeruginosa chemosensory signal transduction complexes localize to cell poles and form and remould in stationary phase. *Mol Microbiol* **61**, 106-118 (2006).

195. Hess, J.F., Bourret, R.B. & Simon, M.I. Histidine phosphorylation and phosphoryl group transfer in bacterial chemotaxis. *Nature* **336**, 139-143 (1988).

196. Mourey, L., Da Re, S., Pedelacq, J.D., Tolstykh, T., Faurie, C., Guillet, V., Stock, J.B. & Samama, J.P. Crystal structure of the CheA histidine phosphotransfer domain that mediates response regulator phosphorylation in bacterial chemotaxis. *J Biol Chem* **276**, 31074-31082 (2001).

197. Quezada, C.M., Gradinaru, C., Simon, M.I., Bilwes, A.M. & Crane, B.R. Helical shifts generate two distinct conformers in the atomic resolution structure of the CheA phosphotransferase domain from Thermotoga maritima. *J Mol Biol* **341**, 1283-1294 (2004).

198. Welch, M., Chinardet, N., Mourey, L., Birck, C. & Samama, J.P. Structure of the CheY-binding domain of histidine kinase CheA in complex with CheY. *Nat Struct Biol* **5**, 25-29 (1998).

199. McEvoy, M.M., Hausrath, A.C., Randolph, G.B., Remington, S.J. & Dahlquist, F.W. Two binding modes reveal flexibility in kinase/response regulator interactions in the bacterial chemotaxis pathway. *Proc Natl Acad Sci USA* **95**, 7333-7338 (1998).

200. Bilwes, A.M., Alex, L.A., Crane, B.R. & Simon, M.I. Structure of CheA, a signal-transducing histidine kinase. *Cell* **96**, 131-41 (1999).

201. Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D. & Yeates, T.O. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proc Natl Acad Sci USA* **96**, 4285-4288 (1999).

202. Vert, J.-P. A tree kernel to analyse phylogenetic profiles. *Bioinformatics* **18**, 276S-284 (2002).

203. Date, S.V. & Marcotte, E.M. Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. *Nat Biotechnol* **21**, 1055-1062 (2003).

204. Ramani, A.K. & Marcotte, E.M. Exploiting the co-evolution of interacting proteins to discover interaction specificity. *J Mol Biol* **327**, 273-284 (2003).

205. Goh, C.S., Bogan, A.A., Joachimiak, M., Walther, D. & Cohen, F.E. Co-evolution of proteins with their interaction partners. *J Mol Biol* **299**, 283-293 (2000).

206. Nei, M. & Kumar, S. *Molecular Evolution and Phylogenetics*, Oxford University Press, New York (2000).

207. Parkhill, J. et al. The genome sequence of the food-borne pathogen Campylobacter jejuni reveals hypervariable sequences. *Nature* **403**, 665-668 (2000).

208. Terry, K., Go, A.C. & Ottemann, K.M. Proteomic mapping of a suppressor of non-chemotactic cheW mutants reveals that Helicobacter pylori contains a new chemotaxis protein. *Mol Microbiol* **61**, 871-882 (2006).

209. Croxen, M.A., Sisson, G., Melano, R. & Hoffman, P.S. The Helicobacter pylori chemotaxis receptor TlpB (HP0103) is required for pH taxis and for colonization of the gastric mucosa. *J Bacteriol* **188**, 2656-2665 (2006).

210. Tomb, J.F. et al. The complete genome sequence of the gastric pathogen Helicobacter pylori. *Nature* **388**, 539-547 (1997).

211. Pittman, M.S., Goodwin, M. & Kelly, D.J. Chemotaxis in the human gastric pathogen Helicobacter pylori: different roles for CheW and the three CheV paralogues, and evidence for CheV2 phosphorylation. *Microbiol* **147**, 2493-504 (2001).

212. Kollmann, M. & Sourjik, V. In silico biology: From simulation to understanding. *Curr Biol* **17**, R132-R134 (2007).

213. Leech, A.J. & Mattick, J.S. Effect of site-specific mutations in different phosphotransfer domains of the chemosensory protein ChpA on Pseudomonas aeruginosa motility. *J Bacteriol* **188**, 8479-8486 (2006).

214. White, O. et al. Genome sequence of the radioresistant bacterium Deinococcus radiodurans R1. *Science* **286**, 1571-1577 (1999).

215. Gupta, R.S. Protein phylogenies and signature sequences: A reappraisal of evolutionary relationships among archaebacteria, eubacteria, and eukaryotes. *Microbiol Mol Biol Rev* **62**, 1435-1491 (1998).

216. Ciccarelli, F.D., Doerks, T., von Mering, C., Creevey, C.J., Snel, B. & Bork, P. Toward automatic reconstruction of a highly resolved tree of life. *Science* **311**, 1283-1287 (2006).

217. Cavalier-Smith, T. A revised six-kingdom system of life. *Biol Rev* **73**, 203-266 (1998).